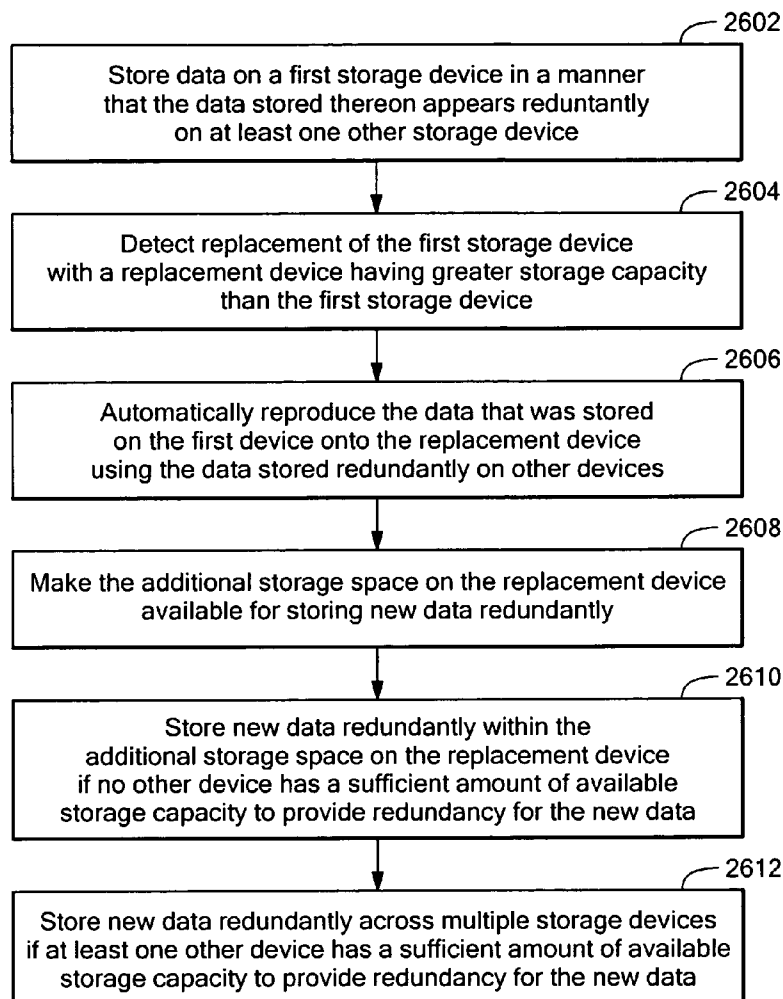




US 20060112222A1

(19) **United States**(12) **Patent Application Publication**
Barrall(10) **Pub. No.: US 2006/0112222 A1**(43) **Pub. Date: May 25, 2006**(54) **DYNAMICALLY EXPANDABLE AND
CONTRACTIBLE FAULT-TOLERANT
STORAGE SYSTEM PERMITTING
VARIOUSLY SIZED STORAGE DEVICES
AND METHOD****Publication Classification**(51) **Int. Cl.**
G06F 12/16 (2006.01)
(52) **U.S. Cl.** 711/114(76) **Inventor: Geoffrey S. Barrall, San Jose, CA (US)**Correspondence Address:
BROMBERG & SUNSTEIN LLP
125 SUMMER STREET
BOSTON, MA 02110-1618 (US)(57) **ABSTRACT**

A dynamically expandable and contractible fault-tolerant storage system permits variously sized storage devices. Data is stored redundantly across one or more storage devices if possible. The layout of data across the one or more storage devices is automatically reconfigured as storage devices are added or removed in order to provide an appropriate level of redundancy for the data to the extent possible. A hash-based compression technique may be used to reduce storage consumption. Techniques for freeing unused storage blocks are also disclosed.

(21) **Appl. No.: 11/267,938**(22) **Filed: Nov. 4, 2005****Related U.S. Application Data**(60) **Provisional application No. 60/625,495, filed on Nov. 5, 2004. Provisional application No. 60/718,768, filed on Sep. 20, 2005.**

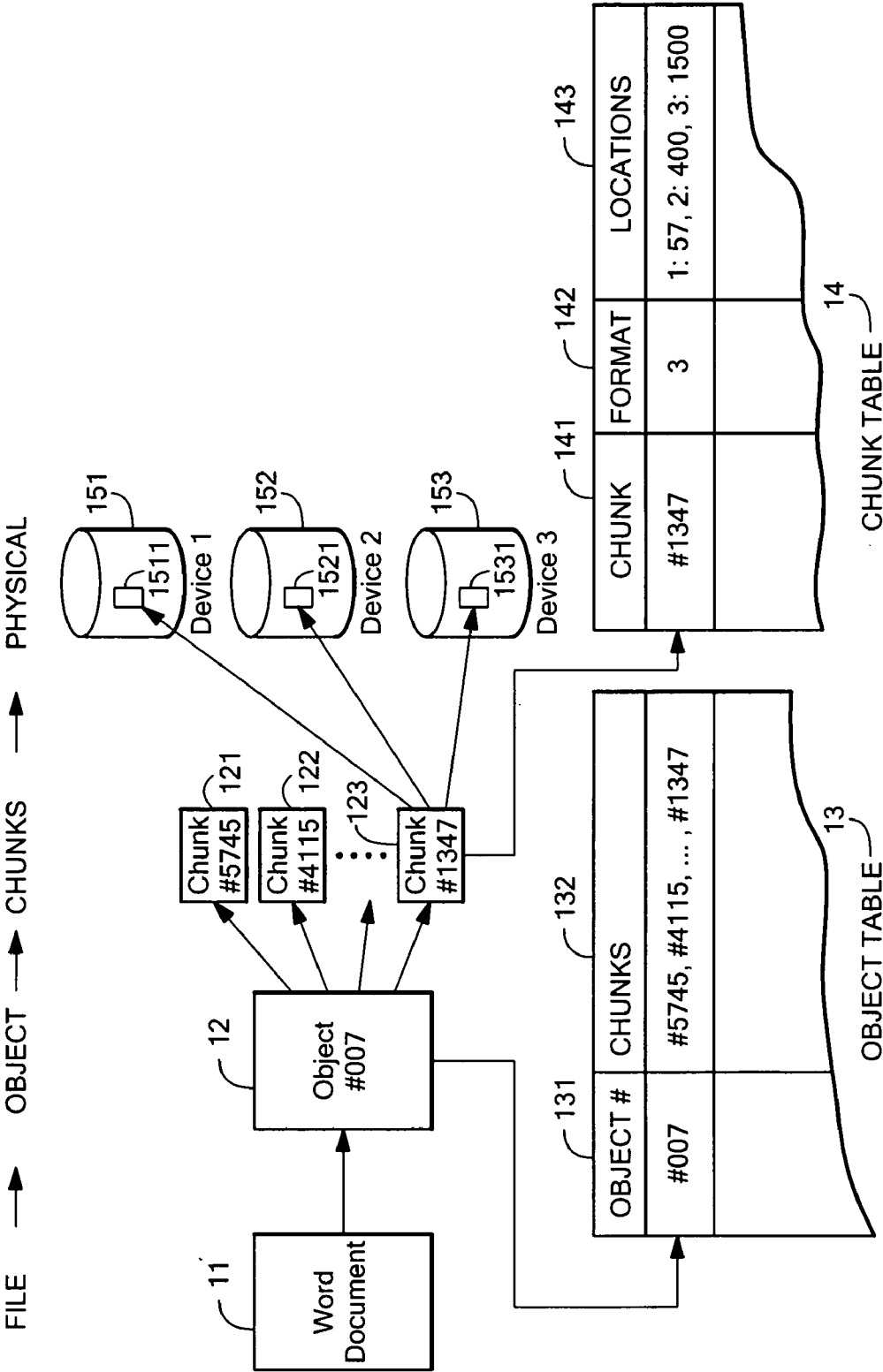


FIG. 1

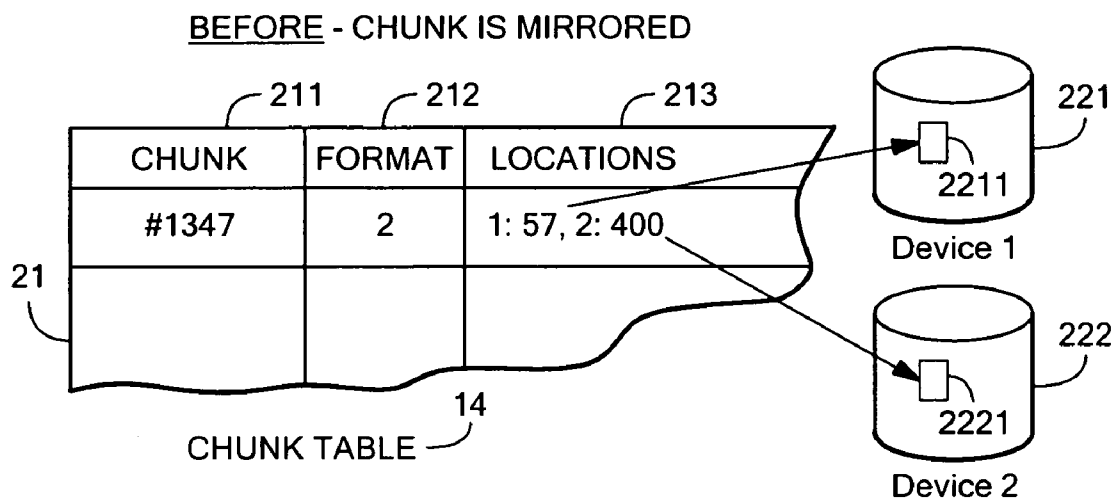


FIG. 2A

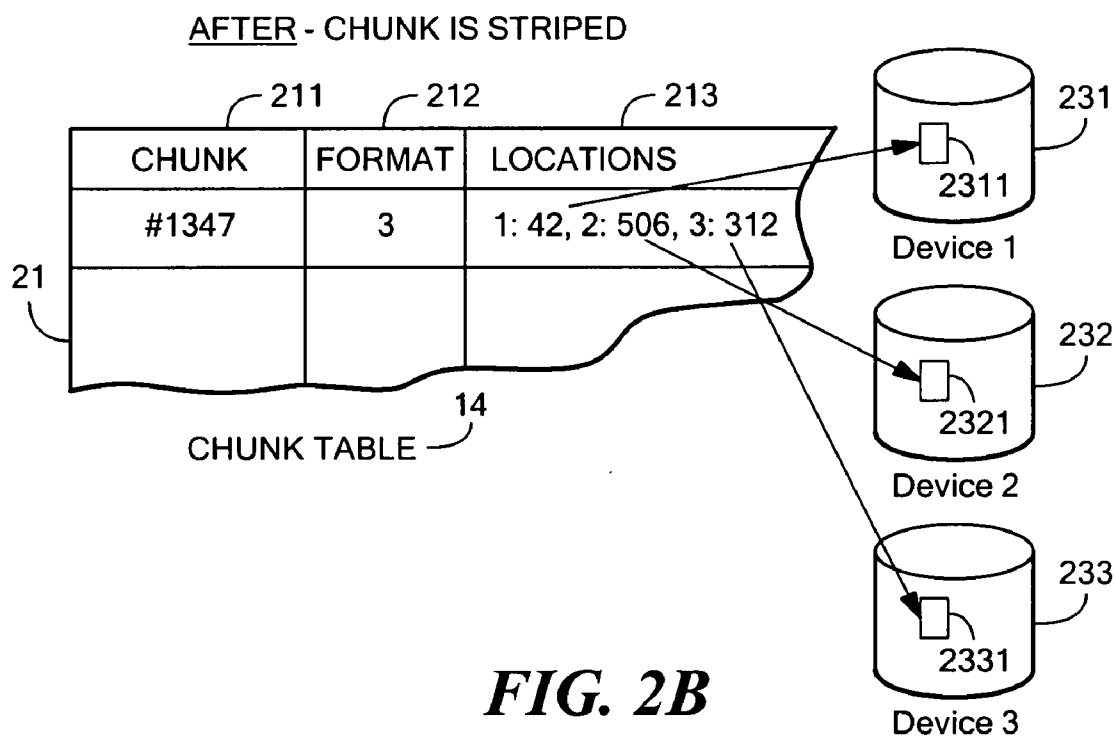
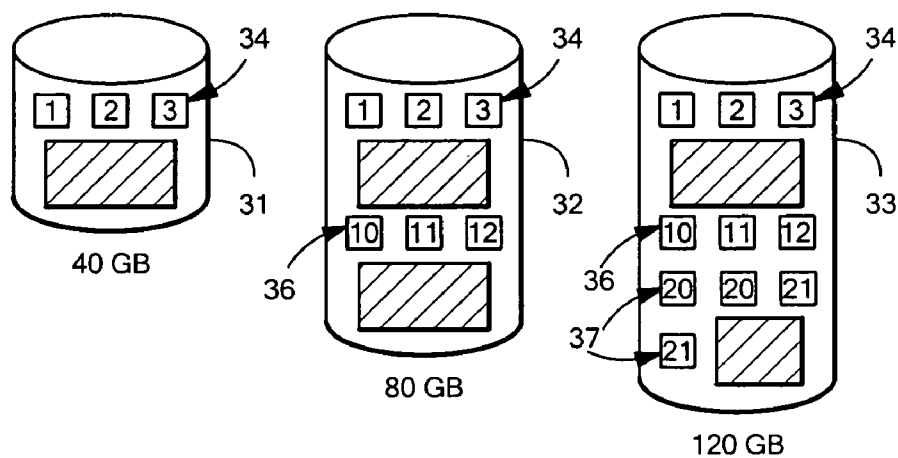


FIG. 2B

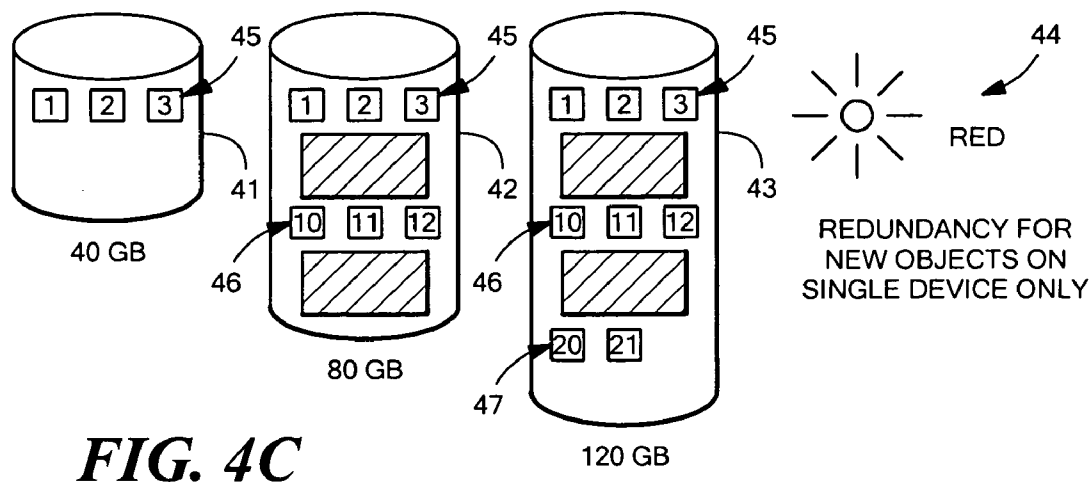
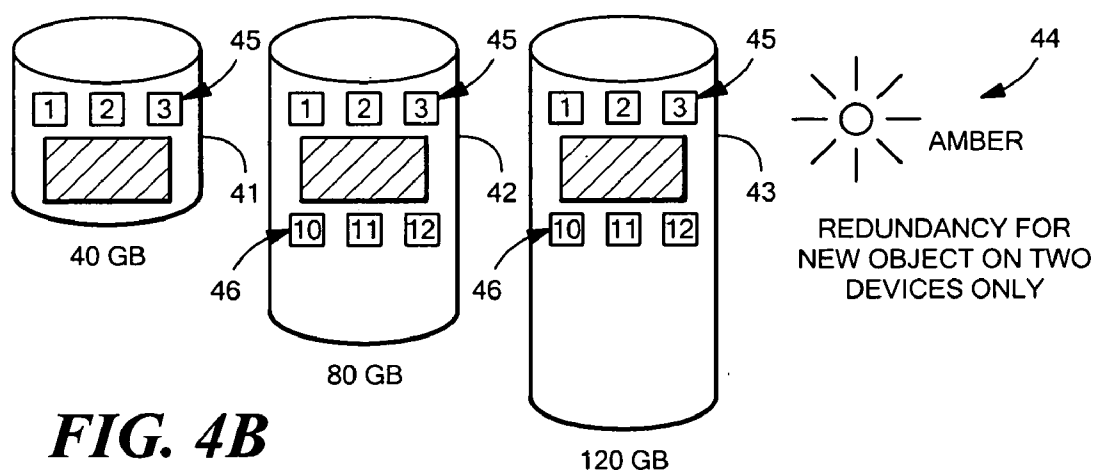
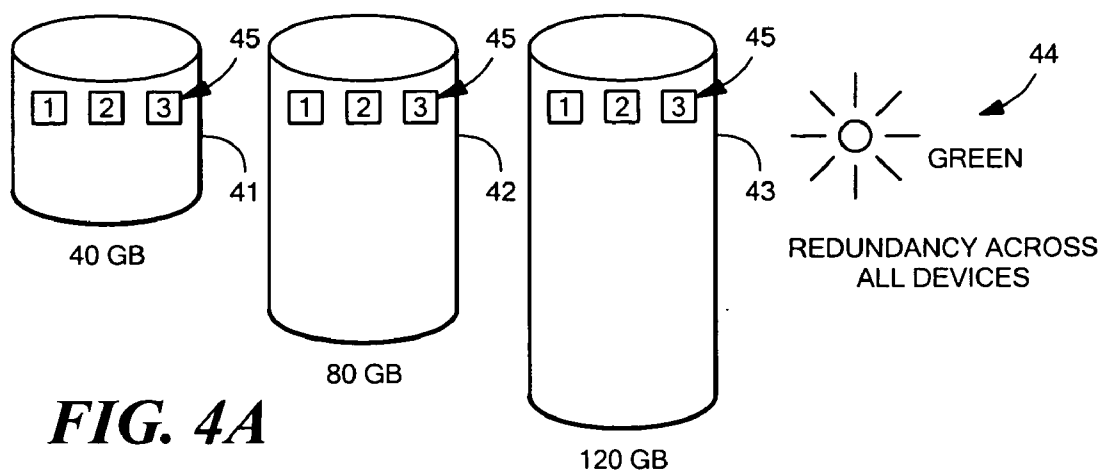


CHUNKS #1, #2, #3 STORED USING PATTERN 3 (STRIPING USING 3 DEVICES)

CHUNKS #10, #11, #12 STORED USING PATTERN 2 (MIRRORED USING 2 DEVICES)

CHUNKS #20, #21 STORED USING PATTERN 1 (MIRRORED USING 1 DEVICE)

FIG. 3



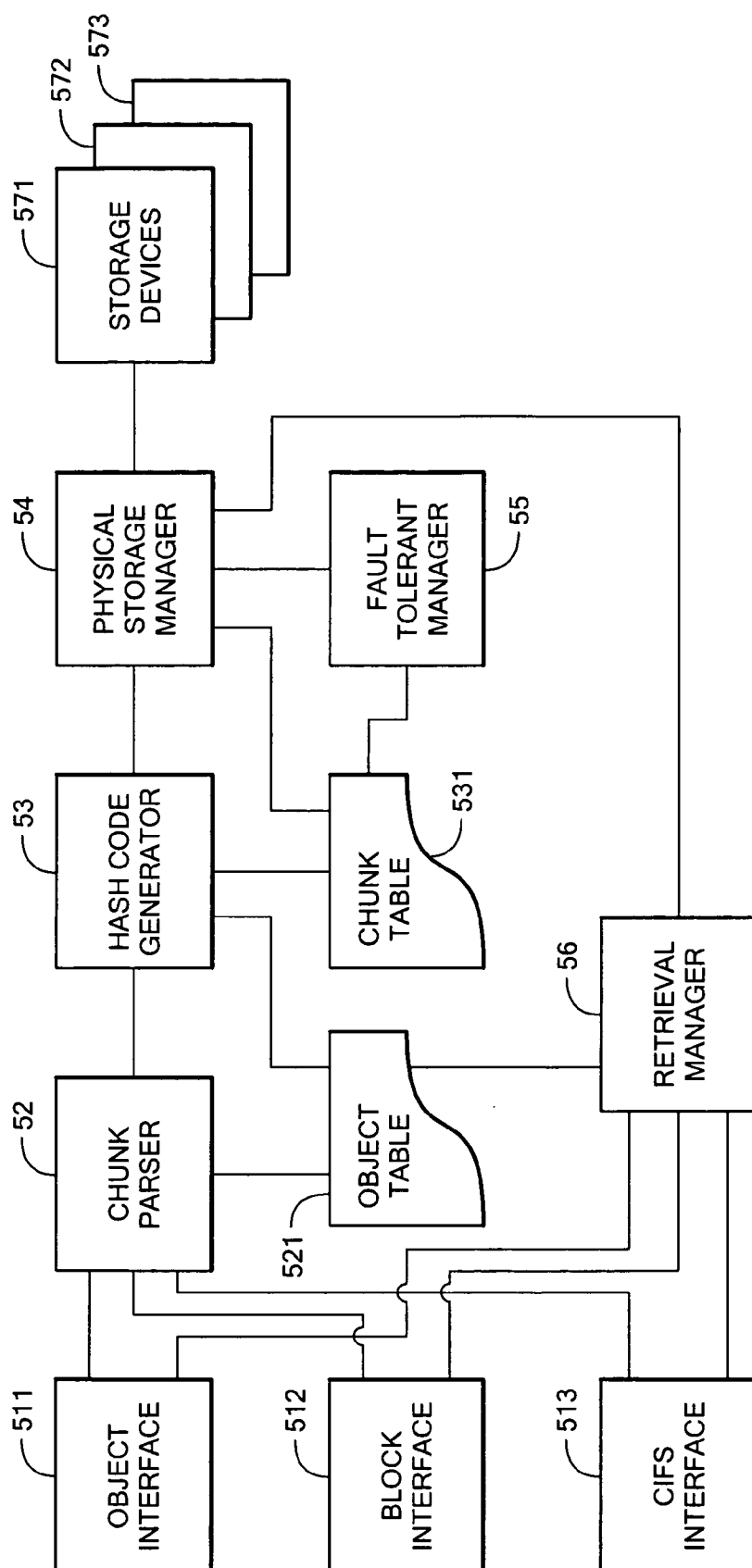


FIG. 5

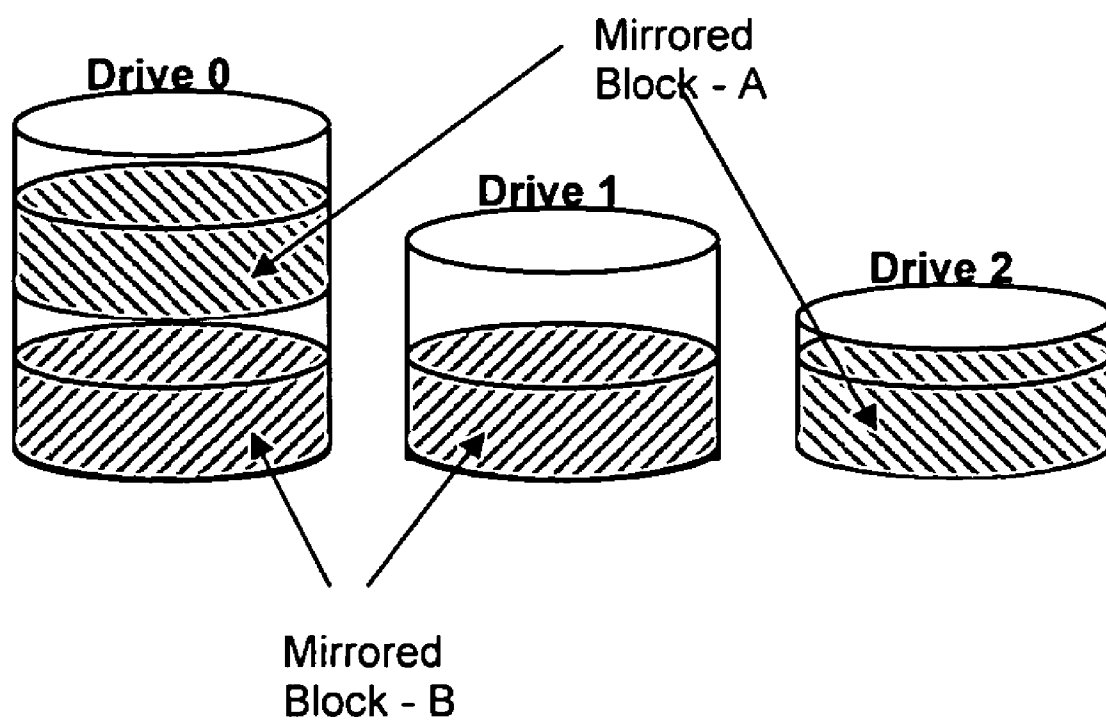


FIG. 6

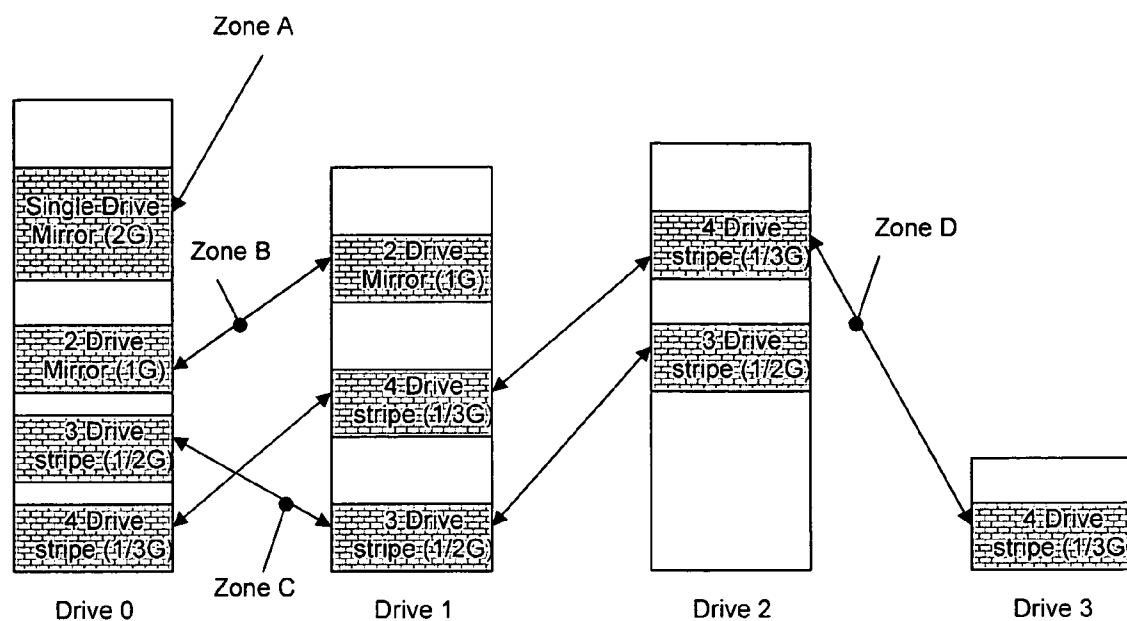


FIG. 7

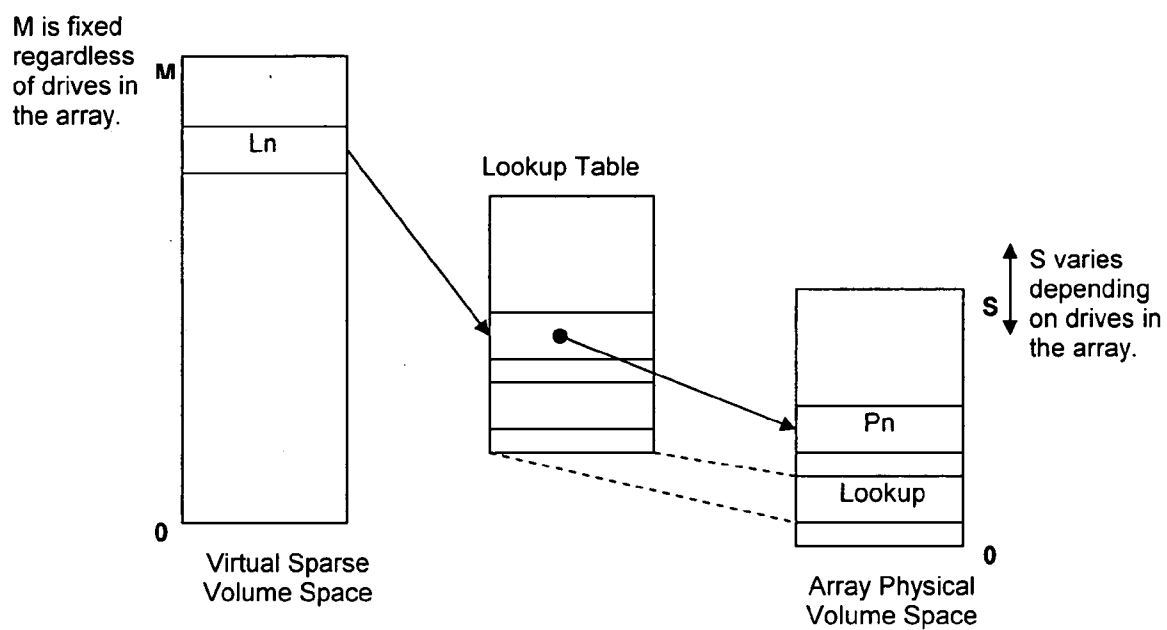
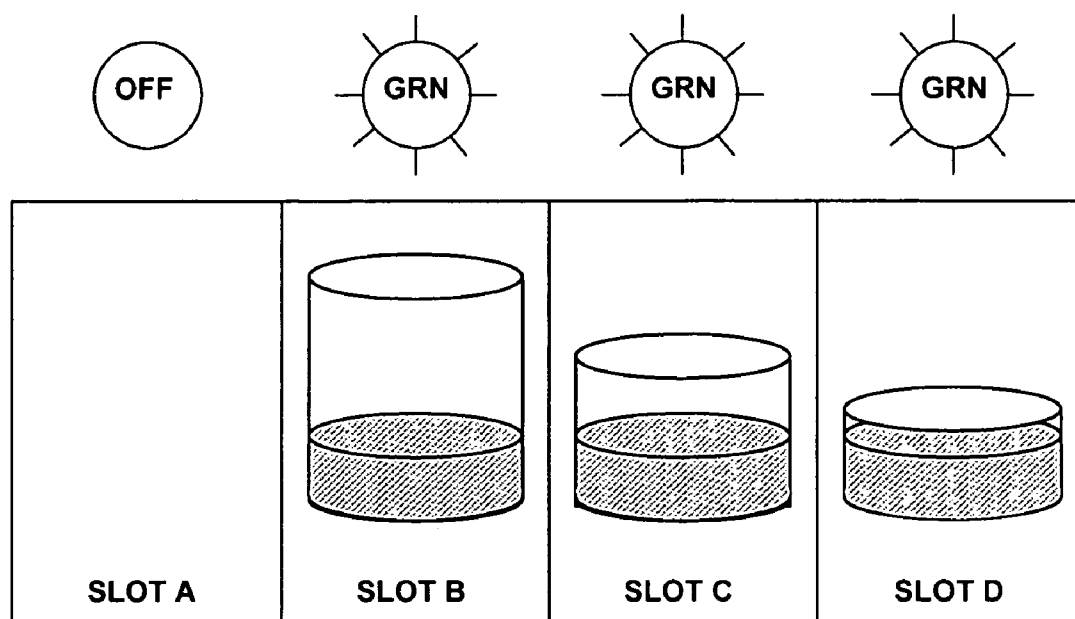
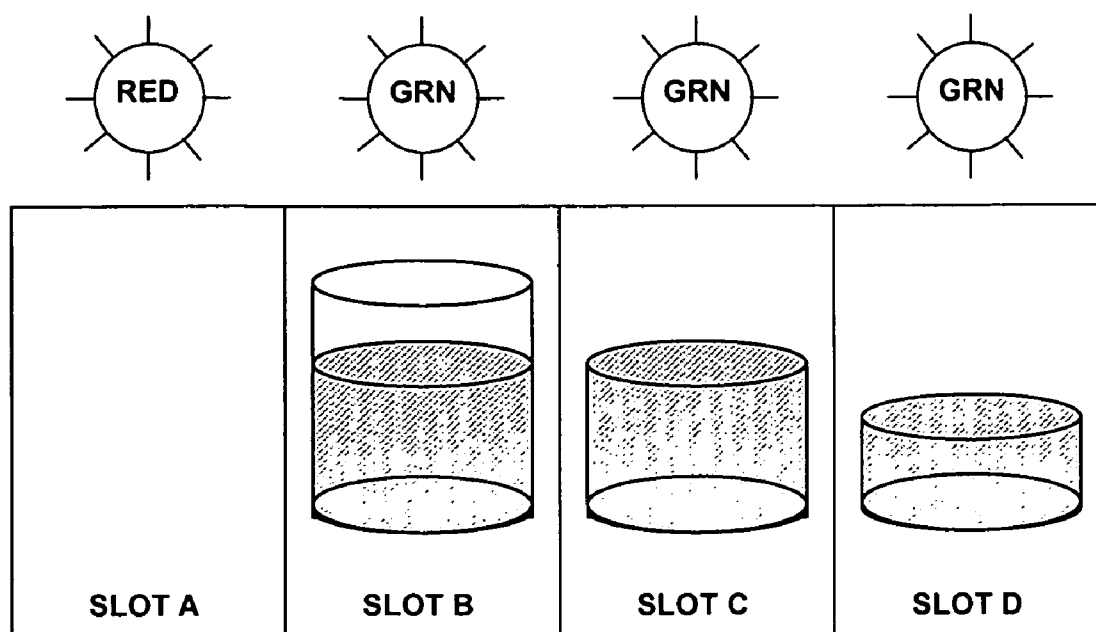


FIG. 8



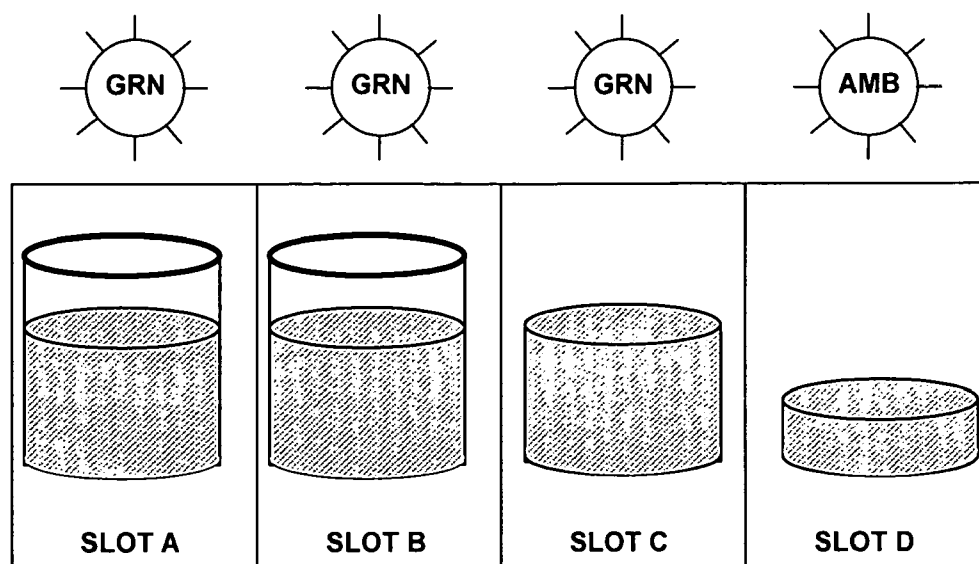
a) All Drives OK with Array space available.

FIG. 9



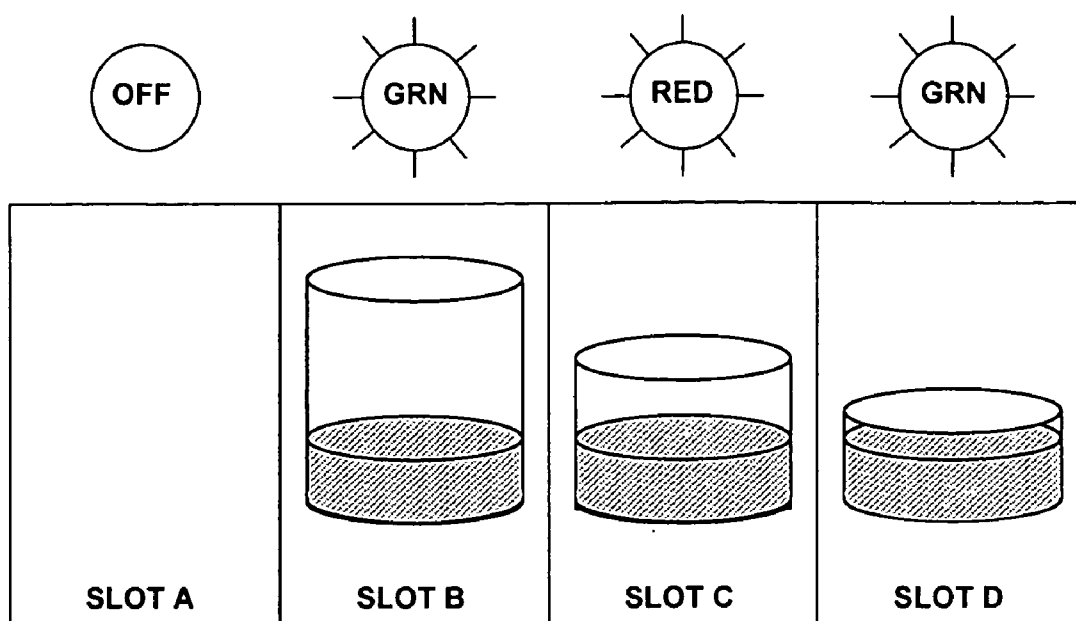
b) Slot C & Slot D full – data not redundant: Add drive to empty Slot A

FIG. 10



c) Array cannot maintain redundant data in the event of a power failure: Replace drive in Slot D

FIG. 11



d) Drive in Slot C has failed

FIG. 12

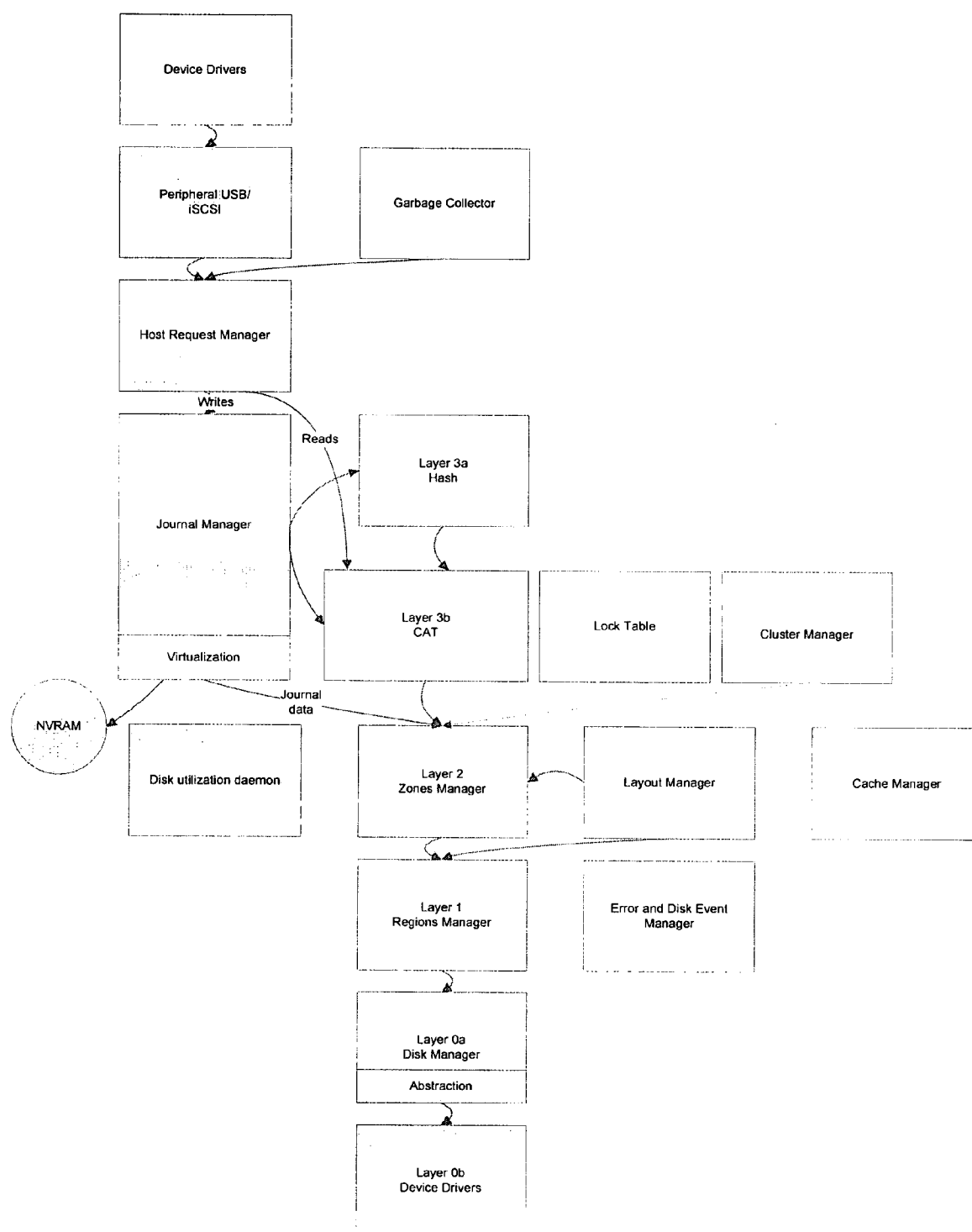


FIG. 13

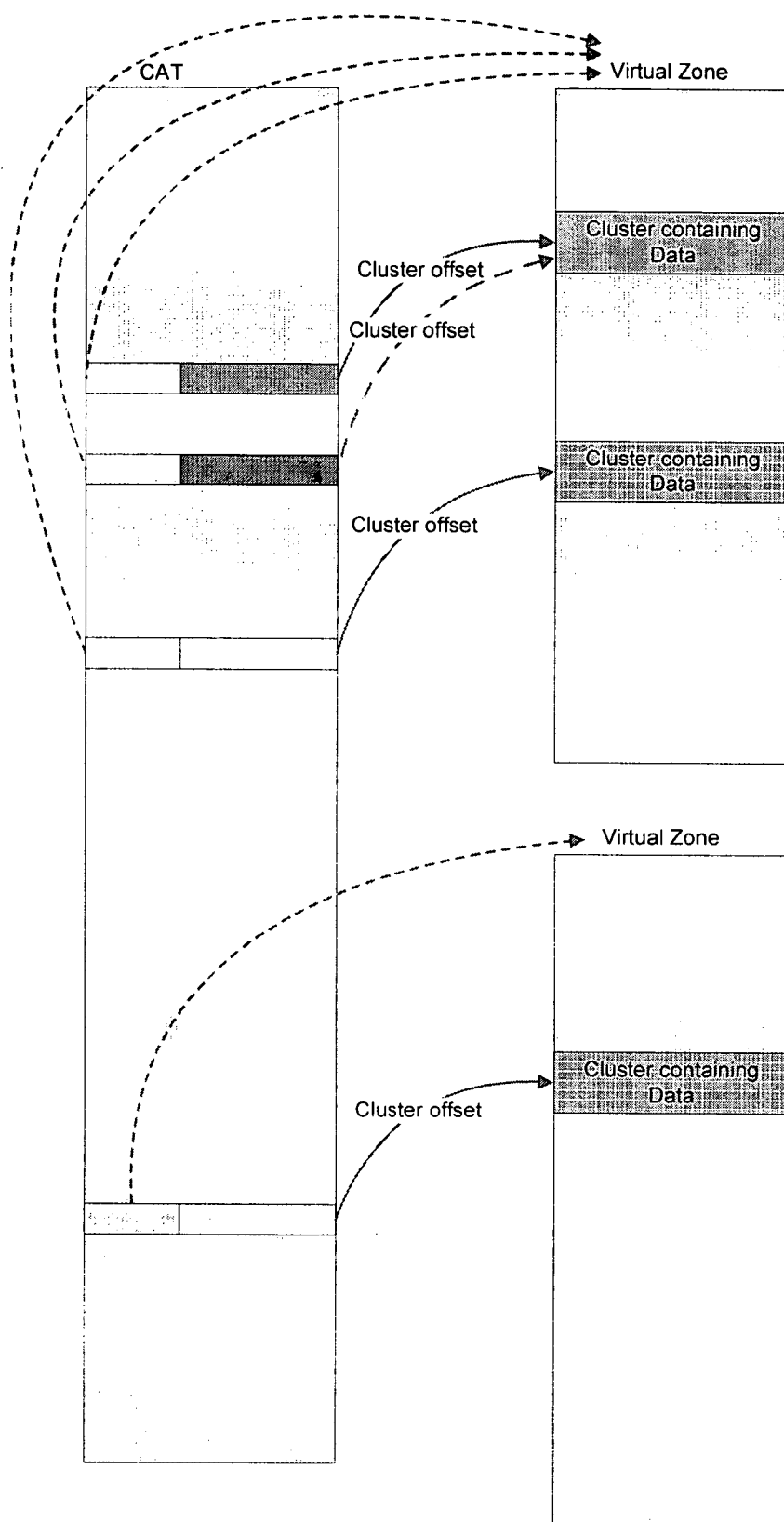


FIG. 14

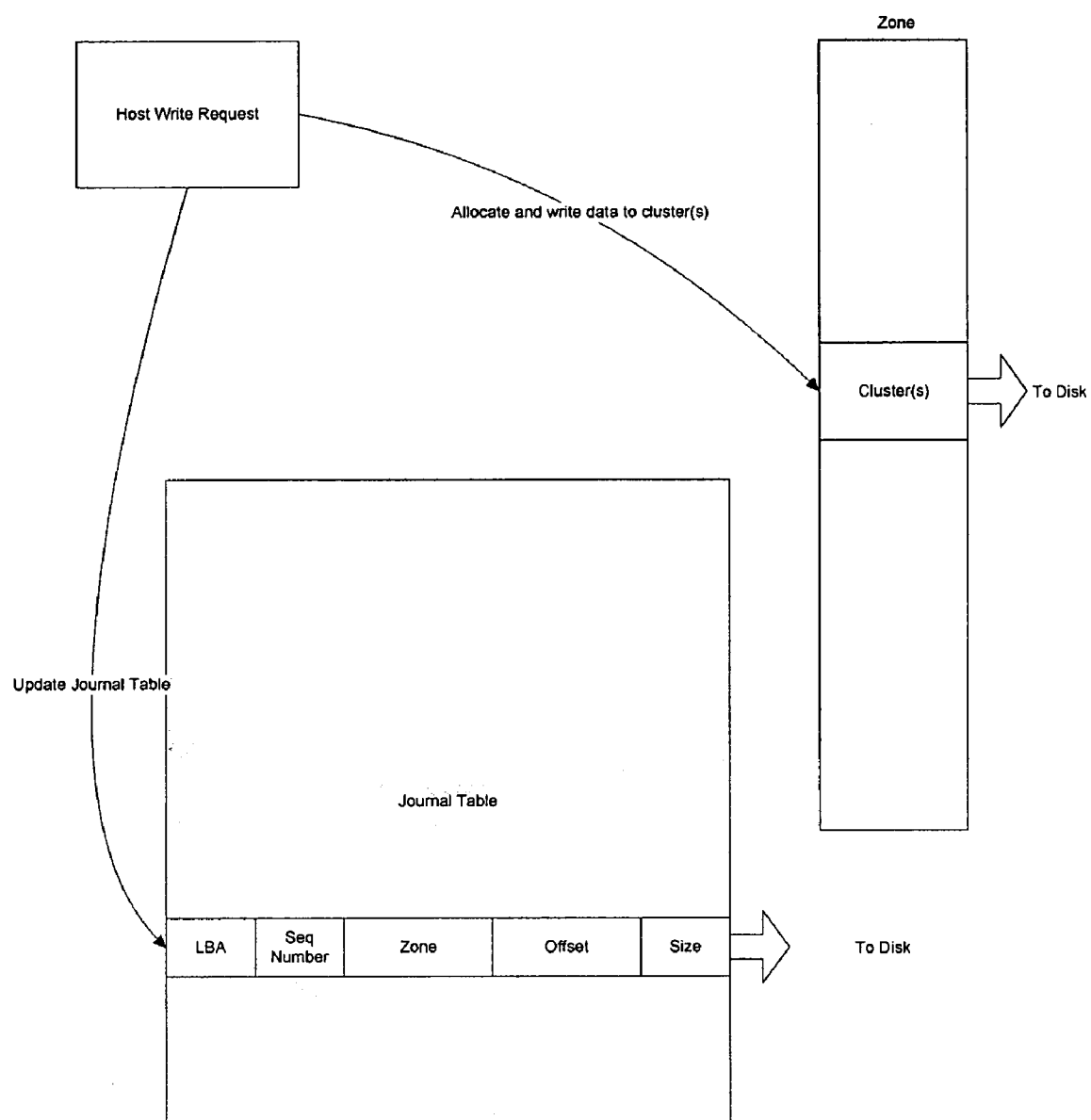


FIG. 15

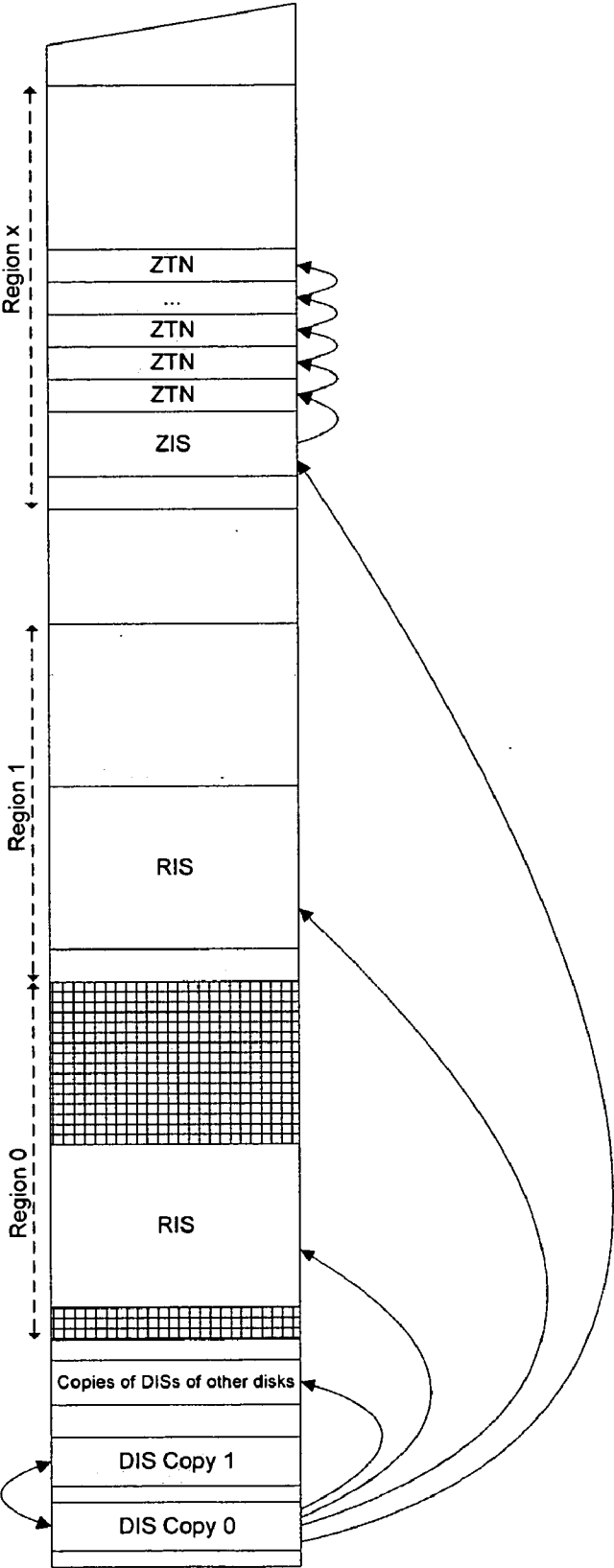


FIG. 16

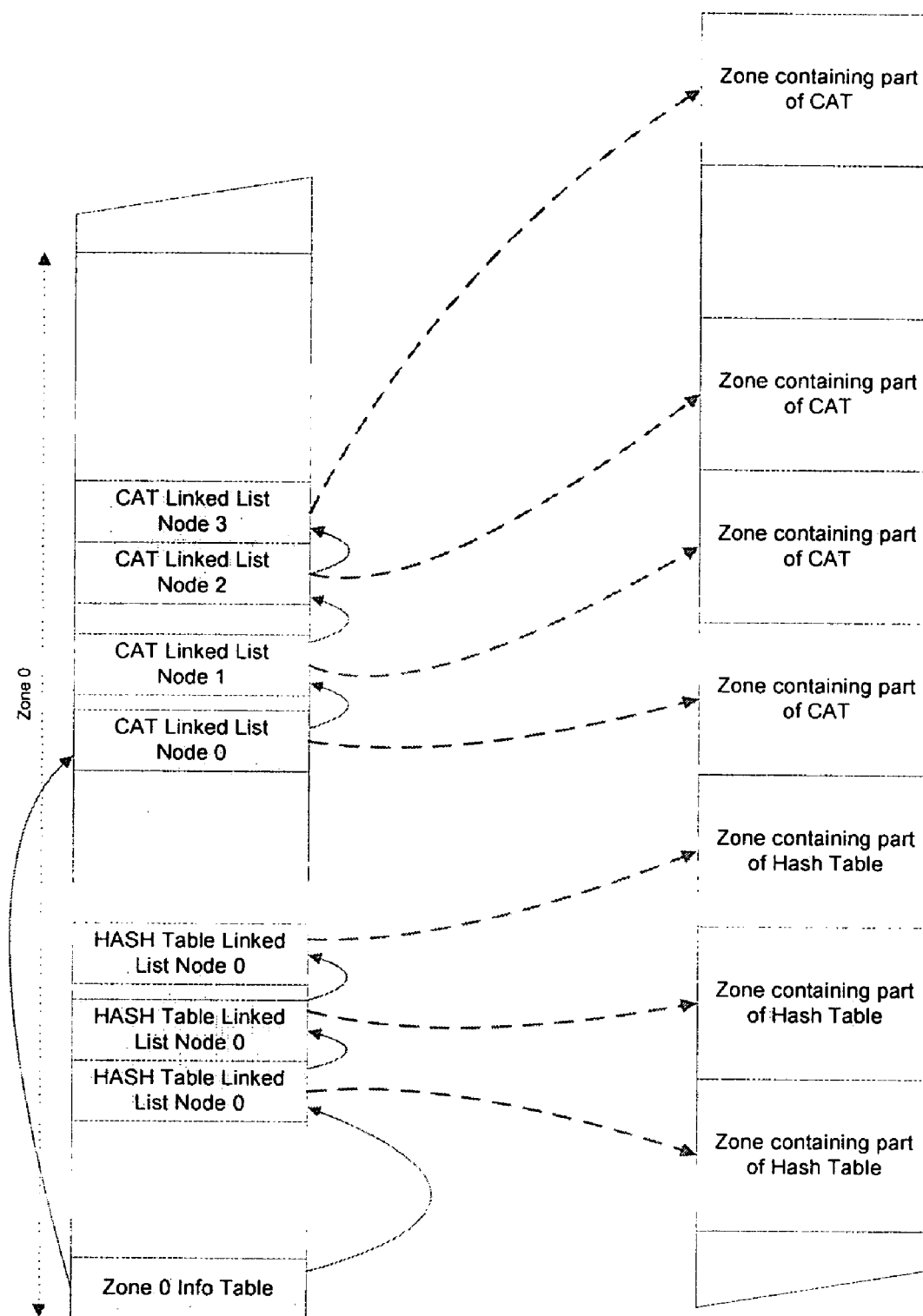


FIG. 17

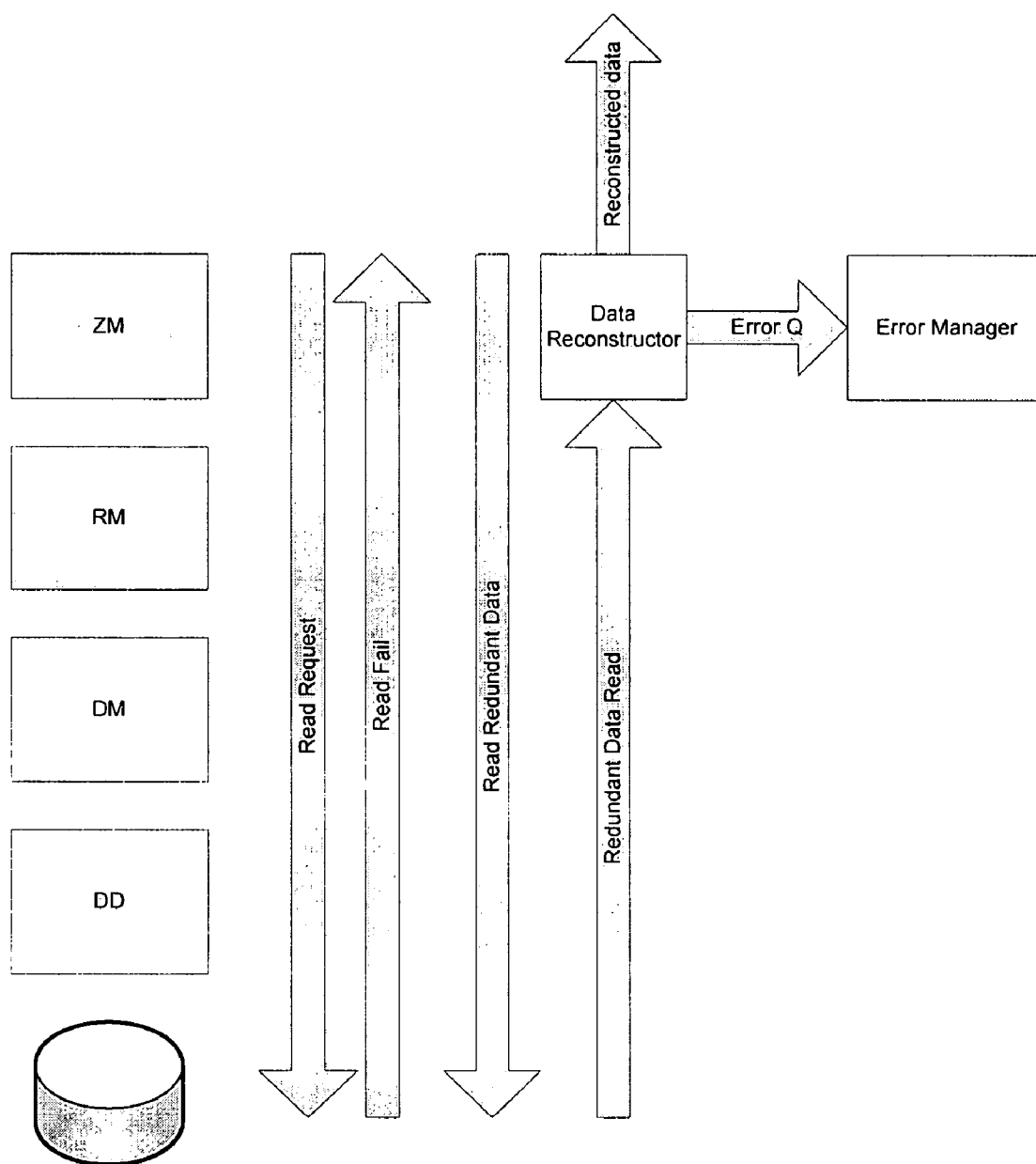


FIG. 18

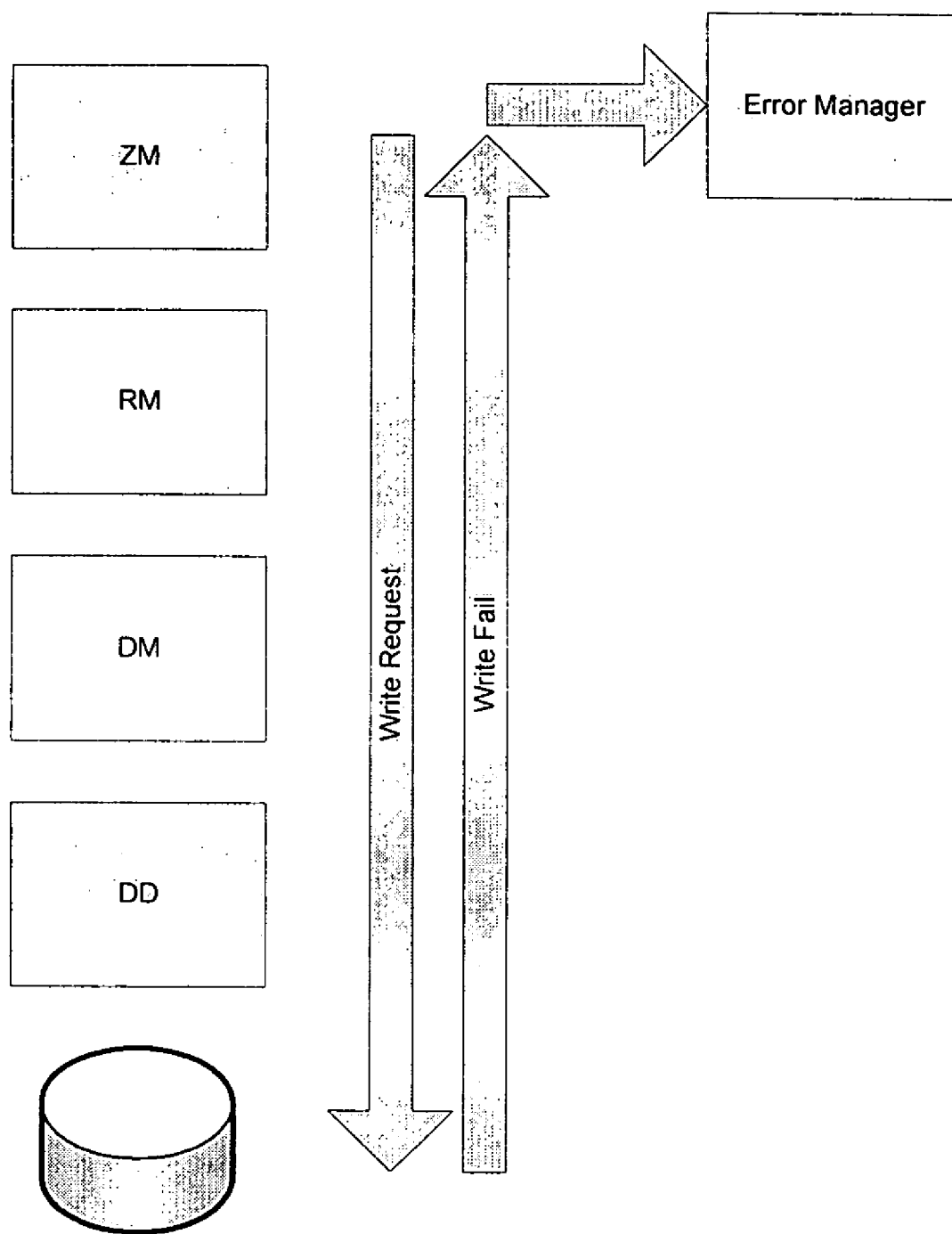


FIG. 19

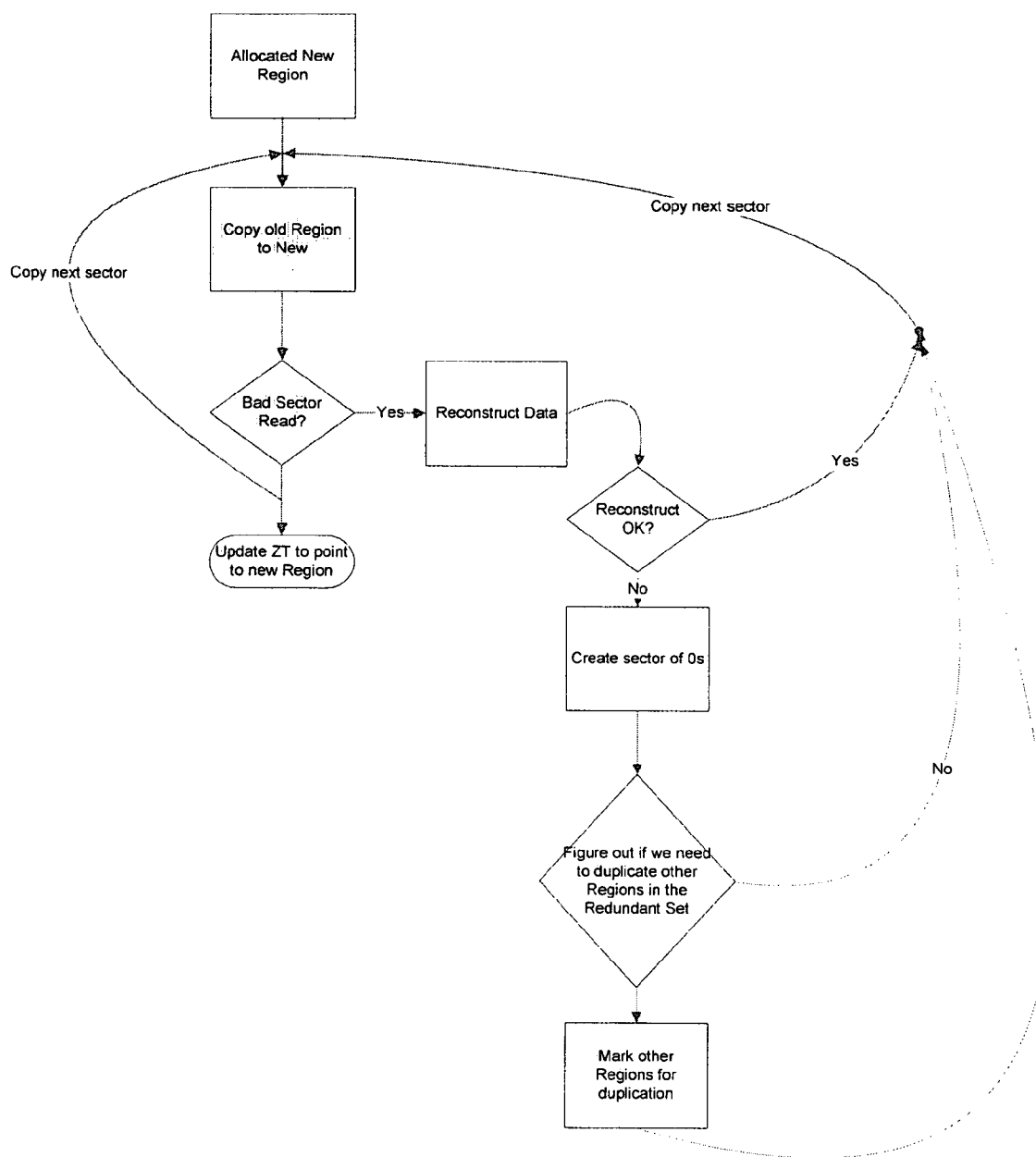


FIG. 20

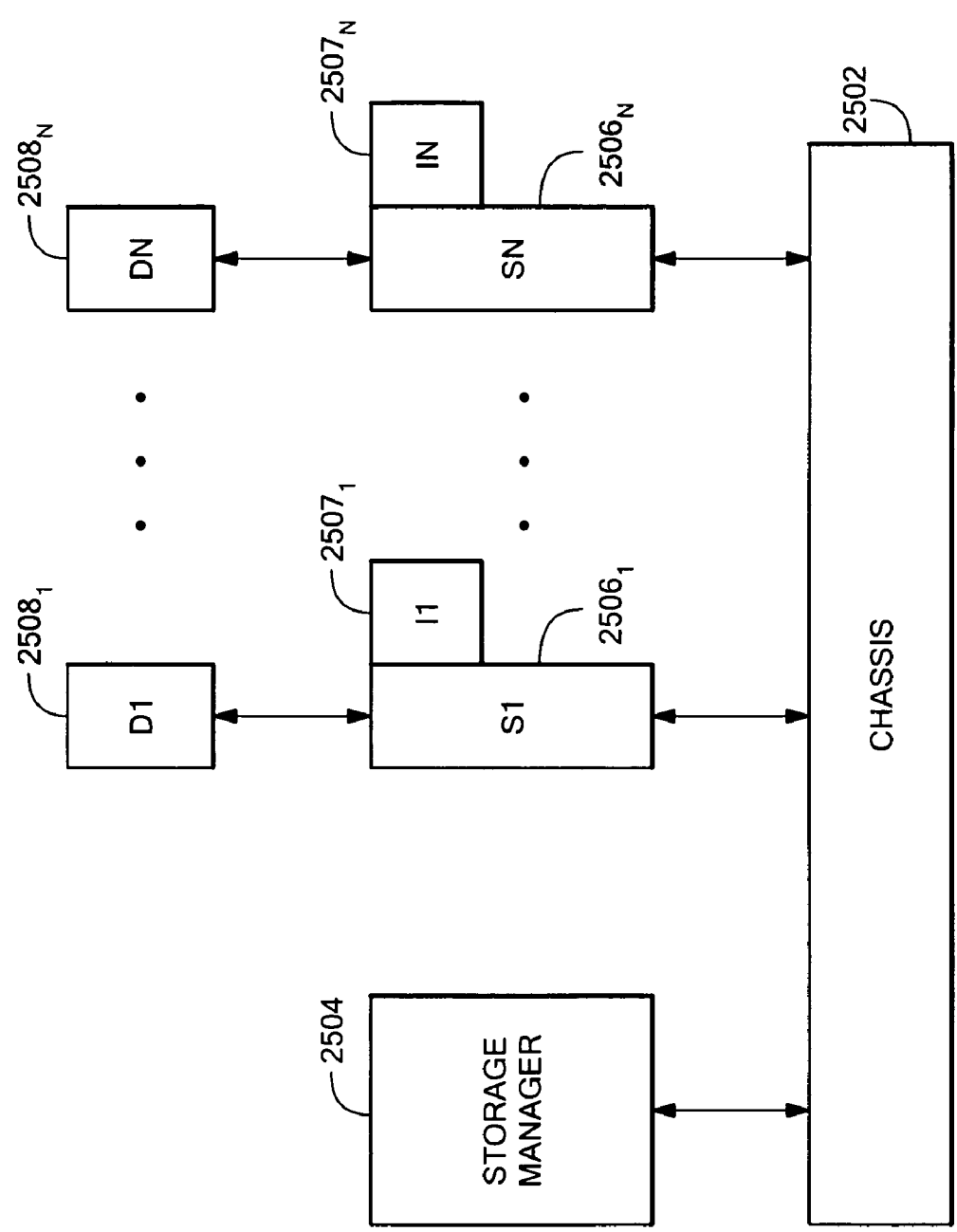


FIG. 21

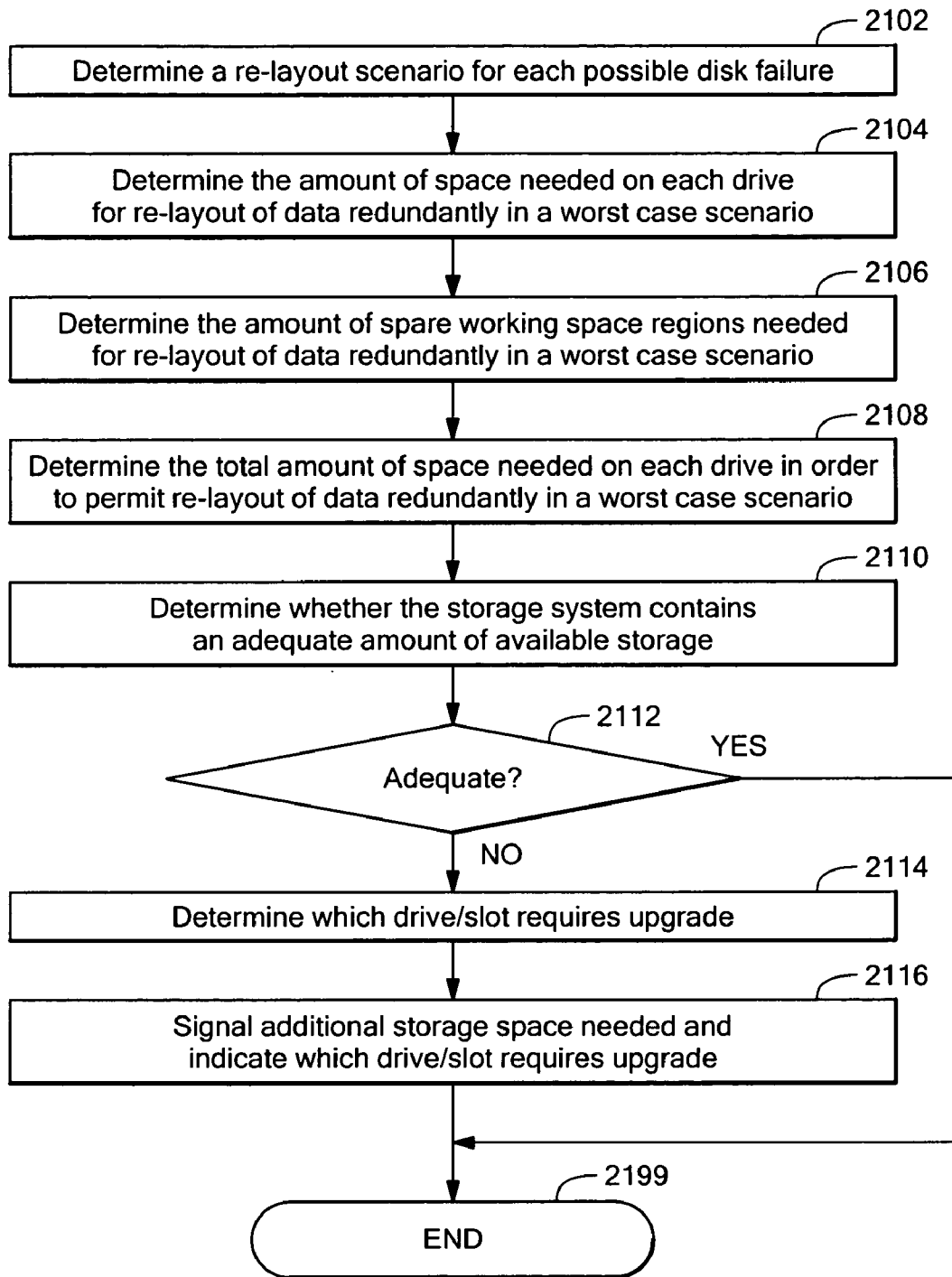


FIG. 22

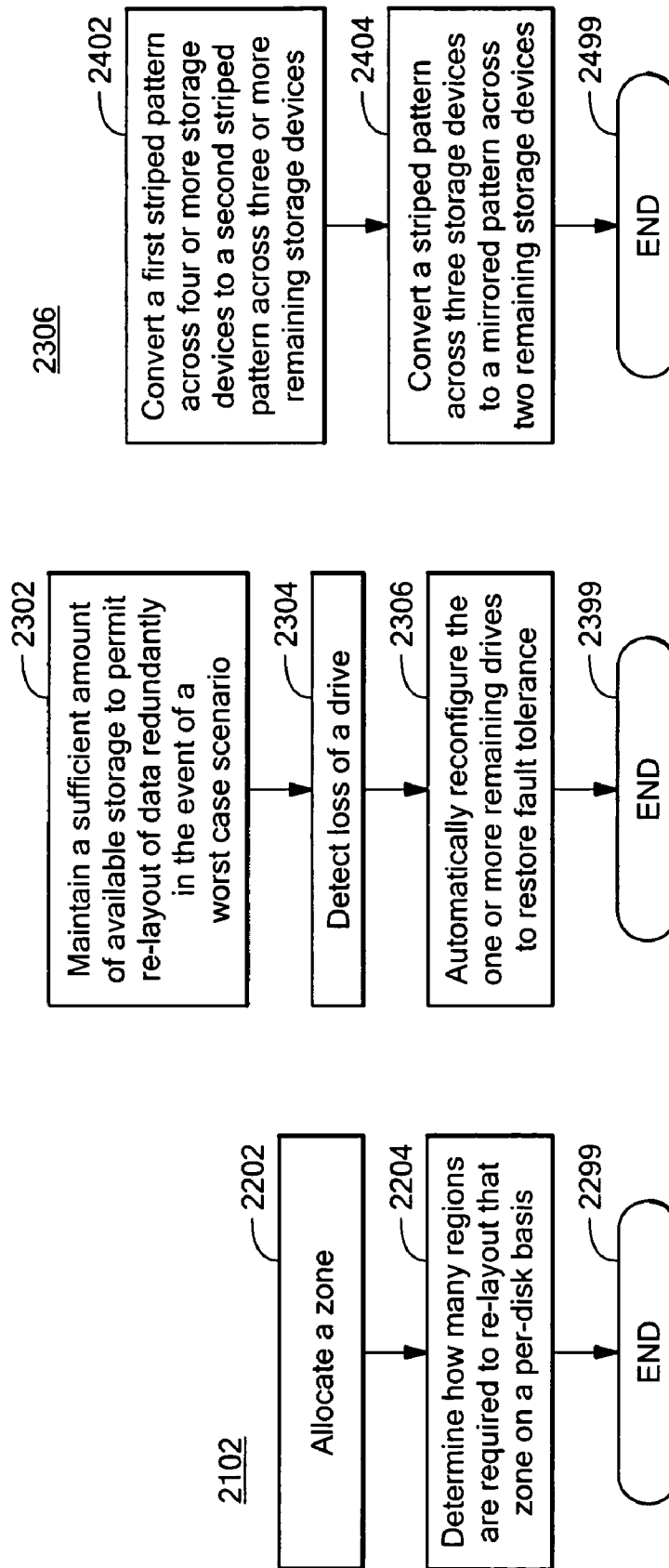
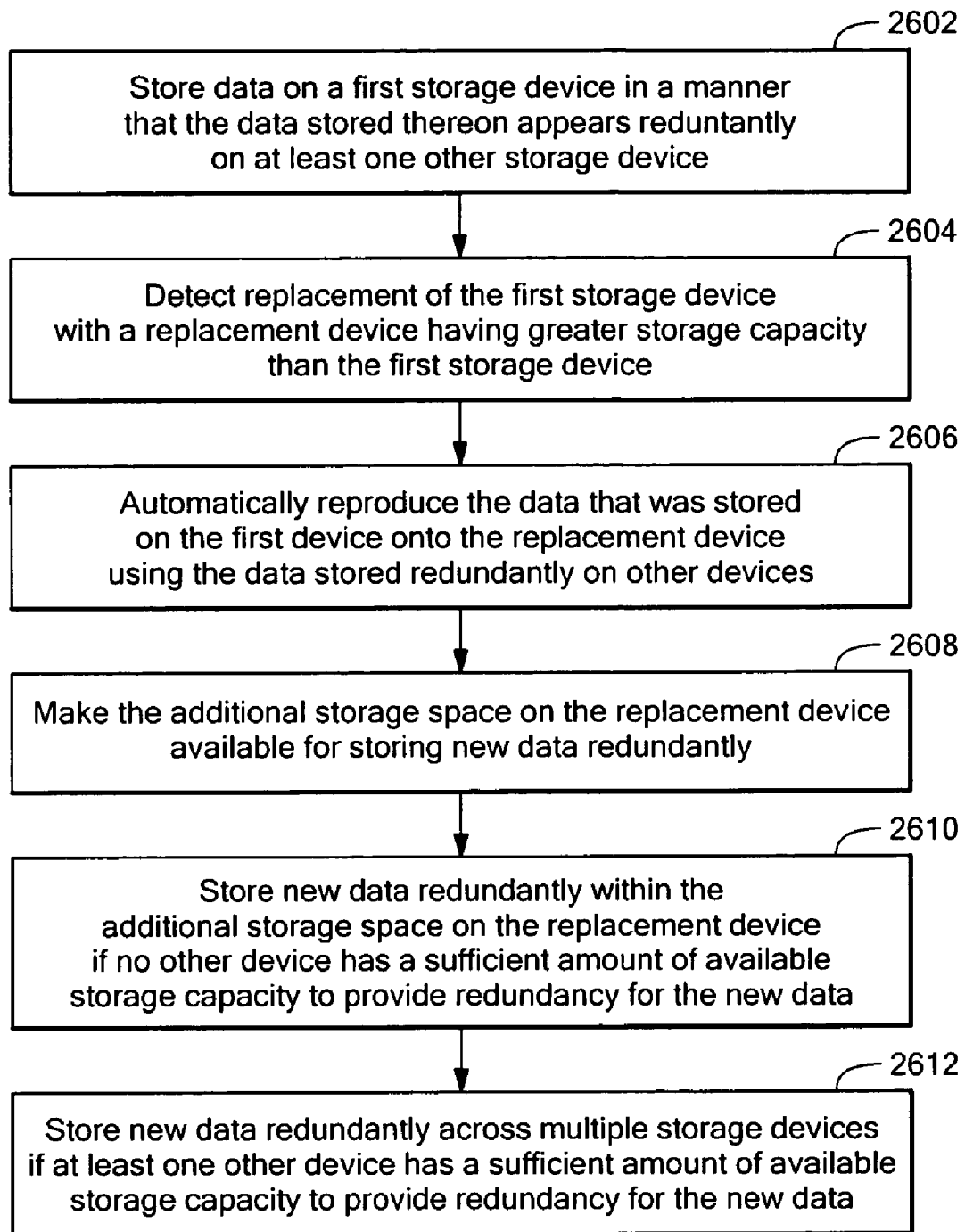


FIG. 23

FIG. 24

FIG. 25

**FIG. 26**

**DYNAMICALLY EXPANDABLE AND
CONTRACTIBLE FAULT-TOLERANT STORAGE
SYSTEM PERMITTING VARIOUSLY SIZED
STORAGE DEVICES AND METHOD**

PRIORITY

[0001] This application claims priority from the following U.S. Provisional Patent Applications, which are hereby incorporated herein by reference in their entireties: 60/625,495 filed on Nov. 5, 2004; and 60/718,768 filed on Sep. 20, 2005.

TECHNICAL FIELD AND BACKGROUND ART

[0002] The present invention relates to digital data storage systems and methods, and more particularly to those providing fault-tolerant storage.

[0003] It is known in the prior art to provide redundant disk storage in a pattern according to any one of various RAID (Redundant Array of Independent Disks) protocols. Typically disk arrays using a RAID pattern are complex structures that require management by experienced information technologists. Moreover in many array designs using a RAID pattern, if the disk drives in the array are of non-uniform capacities, the design may be unable to use any capacity on the drive that exceeds the capacity of the smallest drive in the array.

[0004] One problem with a standard RAID system is that it is possible for disc-surface corruption to occur on an infrequently used area of the disk array. In the event that another drive fails, it is not always possible to determine that corruption has occurred. In such a case, the corrupted data may be propagated and preserved when the RAID array rebuilds the failed drive.

[0005] In many storage systems, a spare storage device will be maintained in a ready state so that it can be used in the event another storage device fails. Such a spare storage device is often referred to as a "hot spare." The hot spare is not used to store data during normal operation of the storage system. When an active storage device fails, the failed storage device is logically replaced by the hot spare, and data is moved or otherwise recreated onto the hot spare. When the failed storage device is repaired or replaced, the data is typically moved or otherwise recreated onto the (re-)activated storage device, and the hot spare is brought offline so that it is ready to be used in the event of another failure. Maintenance of a hot spare disk is generally complex, and so is generally handled by a skilled administrator. A hot spare disk also represents an added expense.

SUMMARY OF THE INVENTION

[0006] In a first embodiment of the invention there is provided a method of storing data in a set of storage devices including at least one storage device. The method involves automatically determining a mixture of redundancy schemes, from among a plurality of redundancy schemes, for storing the data in the set of storage devices in order to provide fault tolerance and, when the set of storage devices happens to include a plurality of devices of different storage capacities, enhanced storage efficiency; and storing the data in the set of storage devices using the mixture of redundancy schemes.

[0007] In a related embodiment, the method may also involve automatically reconfiguring data previously stored on a first arrangement of storage devices using a first redundancy scheme to being stored using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices for accommodating without data loss at least one of expansion of capacity by the addition of another storage device to the set and contraction of capacity by the removal of a storage device from the set.

[0008] In a second embodiment of the invention there is provided a method of storing data in a set of storage devices including at least one storage device in a manner permitting dynamic expansion and contraction of the set. The method involves storing data on the set of storage devices using a first redundancy scheme and, upon a change in the composition of the set of storage devices, automatically reconfiguring the data on the set of storage devices using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices.

[0009] In a third embodiment of the invention there is provided a system for storing data including a set of storage devices having at least one storage device and a storage manager for storing the data in the set of storage devices using a mixture of redundancy schemes. The storage manager automatically determines the mixture of redundancy schemes, from among a plurality of redundancy schemes, for storing the data in the set of storage devices in order to provide fault tolerance and, when the set of storage devices happens to include a plurality of devices of different storage capacities, enhanced storage efficiency.

[0010] In a related embodiment, the storage manager automatically reconfigures data previously stored on a first arrangement of storage devices using a first redundancy scheme to being stored using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices for accommodating without data loss at least one of expansion of capacity by the addition of another storage device to the set and contraction of capacity by the removal of a storage device from the set.

[0011] In a fourth embodiment of the invention there is provided a system for storing data including a set of storage devices having at least one storage device and a storage manager for storing the data in the set of storage devices. The storage manager stores data on the set of storage devices using a first redundancy scheme and, upon a change in the composition of the set of storage devices, automatically reconfigures the data on the set of storage devices using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices.

[0012] In related embodiments, the redundancy schemes may be selected from the group including mirroring, striping

with parity, RAID6, dual parity, diagonal Parity, Low Density Parity Check codes, and turbo codes.

[0013] In other related embodiments, the data may be reconfigured by converting a mirrored pattern across two storage devices to a mirrored pattern on a single storage device; converting a mirrored pattern across a first pair of storage devices to a mirrored pattern across a second pair of storage devices; converting a striped pattern across three storage devices to a mirrored pattern across two storage devices; converting a first striped pattern across a first plurality of storage devices to a second striped pattern across a second plurality of storage devices; converting a mirrored pattern across one storage device to a mirrored pattern across two storage devices; converting a mirrored pattern on one storage device to a mirrored pattern on another storage device; converting a mirrored pattern across two storage devices to striped pattern across three storage devices; or converting a first striped pattern across a first plurality of storage devices to a second striped pattern across a second plurality of storage devices including a storage device added to the set.

[0014] In still other related embodiments, an object may be parsed into chunks, each chunk having particular content, and only chunks having unique content may be stored in the storage system. The chunks may be identified as having unique content based on a predetermined hash function. In such embodiments, storing only chunks identified as being unique may involve computing a hash number for each chunk and storing in the storage system a unique identifier for the object linked to hash numbers for the stored chunks and location information therefor. Storing a unique identifier for the object may involve using an object table that lists the unique identifier for the object and the hash numbers for the stored chunks. Storing location information may involve storing in a chunk table a listing for each chunk stored therein including a representation of its physical location in the storage system. Objects stored in the storage system may be translated into a representation of at least one of a single storage device and a file system and an object storage system. Physical allocation of storage of chunks in the storage system may be managed so as to provide fault-tolerant storage, even when the storage system may include storage devices of varying capacity. Upon retrieving a particular chunk from the storage system, the contents of the retrieved chunk may be verified based on the stored hash number for the chunk, for example, by computing a hash number for the retrieved chunk and comparing the computed hash number with the corresponding stored hash number for the chunk.

[0015] In another embodiment of the invention there is provided a method of storing data in a set of storage devices involving storing data on the set of storage devices using at least one redundancy scheme, detecting removal of a storage device from the set, reconfiguring a portion of the data from the removed storage device using the remaining storage devices, detecting reinsertion of the removed storage device into the set, freeing regions of the reinserted storage device corresponding to said portion, and refreshing any remaining regions of the reinserted storage device that may have been modified between removal and reinsertion.

[0016] In yet another embodiment of the invention there is provided a method for freeing unused storage blocks in a

storage system. The method involves identifying a free block, determining a logical sector address for the free block, identifying a cluster access table entry for the free block based on the logical sector address, decrementing a reference count associated with the cluster access table entry, and freeing the block if the decremented reference count is zero.

[0017] In related embodiments, identifying a free block may involve identifying a free block bitmap associated with the host filesystem and parsing the free block bitmap to identify clusters that are no longer being used by the filesystem. Identifying the free block bitmap associated with the host filesystem may involve maintaining a partition table, parsing the partition table to locate an operating system partition, parsing the operating system partition to locate a superblock, and identifying the host filesystem based on the superblock. A working copy of the free block bitmap may be made, in which case the working copy may be parsed to identify clusters that are no longer being used by the filesystem.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] The foregoing features of the invention will be more readily understood by reference to the following detailed description, taken with reference to the accompanying drawings, in which:

[0019] **FIG. 1** is an illustration of an embodiment of the invention in which an object is parsed into a series of chunks for storage.

[0020] **FIG. 2** illustrates in the same embodiment how a pattern for fault-tolerant storage for a chunk may be dynamically changed as a result of the addition of more storage.

[0021] **FIG. 3** illustrates in a further embodiment of the invention the storage of chunks in differing fault-tolerant patterns on a storage system constructed using different sized storage devices.

[0022] **FIG. 4** illustrates another embodiment of the invention in which indicator states are used to warn of inefficient storage use and low levels of fault tolerance.

[0023] **FIG. 5** is a block diagram of functional modules used in the storage, retrieval and re-layout of data in accordance with an embodiment of the invention.

[0024] **FIG. 6** shows an example in which mirroring is used in an array containing more than two drives.

[0025] **FIG. 7** shows some exemplary zones using different layout schemes to store their data.

[0026] **FIG. 8** shows a lookup table for implementing sparse volumes.

[0027] **FIG. 9** shows status indicators for an exemplary array having available storage space and operating in a fault-tolerant manner, in accordance with an exemplary embodiment of the present invention.

[0028] **FIG. 10** shows status indicators for an exemplary array that does not have enough space to maintain redundant data storage and more space must be added, in accordance with an exemplary embodiment of the present invention.

[0029] **FIG. 11** shows status indicators for an exemplary array that would be unable to maintain redundant data in the

event of a failure, in accordance with an exemplary embodiment of the present invention.

[0030] **FIG. 12** shows status indicators for an exemplary array in which a storage device has failed, in accordance with an exemplary embodiment of the present invention. Slots B, C, and D are populated with storage devices.

[0031] **FIG. 13** shows a module hierarchy representing the different software layers of an exemplary embodiment and how they relate to one another.

[0032] **FIG. 14** shows how a cluster access table is used to access a data clusters in a Zone, in accordance with an exemplary embodiment of the present invention.

[0033] **FIG. 15** shows a journal table update in accordance with an exemplary embodiment of the present invention.

[0034] **FIG. 16** shows drive layout in accordance with an exemplary embodiment of the invention.

[0035] **FIG. 17** demonstrates the layout of Zone 0 and how other zones are referenced, in accordance with an exemplary embodiment of the invention.

[0036] **FIG. 18** demonstrates read error handling in accordance with an exemplary embodiment of the invention.

[0037] **FIG. 19** demonstrates write error handling in accordance with an exemplary embodiment of the invention.

[0038] **FIG. 20** is a logic flow diagram demonstrating backup of a bad Region by the Error Manager in accordance with an exemplary embodiment of the invention.

[0039] **FIG. 21** is a schematic block diagram showing the relevant components of a storage array in accordance with an exemplary embodiment of the present invention.

[0040] **FIG. 22** is a logic flow diagram showing exemplary logic for managing a virtual hot spare in accordance with an exemplary embodiment of the present invention.

[0041] **FIG. 23** is a logic flow diagram showing exemplary logic for determining a re-layout scenario for each possible disk failure, as in block 2102 of **FIG. 22**, in accordance with an exemplary embodiment of the present invention.

[0042] **FIG. 24** is a logic flow diagram showing exemplary logic for invoking the virtual hot spare functionality in accordance with an exemplary embodiment of the present invention.

[0043] **FIG. 25** is a logic flow diagram showing exemplary logic for automatically reconfiguring the one or more remaining drives to restore fault tolerance for the data, as in block 2306 of **FIG. 24**, in accordance with an exemplary embodiment of the present invention.

[0044] **FIG. 26** is a logic flow diagram showing exemplary logic for upgrading a storage device, in accordance with an exemplary embodiment of the present invention.

DETAILED DESCRIPTION OF SPECIFIC EMBODIMENTS

[0045] Definitions. As used in this description and the accompanying claims, the following terms shall have the meanings indicated, unless the context otherwise requires:

[0046] A “chunk” of an object is an abstract slice of an object, made independently of any physical storage being used, and is typically a fixed number of contiguous bytes of the object.

[0047] A fault-tolerant “pattern” for data storage is the particular which by data is distributed redundantly over one or more storage devices, and may be, among other things, mirroring (e.g., in a manner analogous to RAID1), striping (e.g., in a manner analogous to RAID5), RAID6, dual parity, diagonal Parity, Low Density Parity Check codes, turbo codes, or other redundancy scheme or combination of redundancy schemes.

[0048] A hash number for a given chunk is “unique” when the given chunk produces a hash number that generally will differ from the hash number for any other chunk, except when the other chunk has data content identical to the given chunk. That is, two chunks will generally have different hash numbers whenever their content is non-identical. As described in further detail below, the term “unique” is used in this context to cover a hash number that is generated from a hash function occasionally producing a common hash number for chunks that are non-identical because hash functions are not generally perfect at producing different numbers for different chunks.

[0049] A “Region” is a set of contiguous physical blocks on a storage medium (e.g., hard drive).

[0050] A “Zone” is composed of two or more Regions. The Regions that make up a Zone are generally not required to be contiguous. In an exemplary embodiment as described below, a Zone stores the equivalent of 1 GB of data or control information.

[0051] A “Cluster” is the unit size within Zones and represents a unit of compression (discussed below). In an exemplary embodiment as described below, a Cluster is 4 KB (i.e., eight 512-byte sectors) and essentially equates to a Chunk.

[0052] A “Redundant set” is a set of sectors/clusters that provides redundancy for a set of data.

[0053] “Backing up a Region” involves copying the contents of one Region to another Region.

[0054] A “first pair” and a “second pair” of storage devices may include a common storage device.

[0055] A “first plurality” and a “second plurality” of storage devices may include one or more common storage devices.

[0056] A “first arrangement” and a “second arrangement” or “different arrangement” of storage devices may include one or more common storage devices.

[0057] **FIG. 1** is an illustration of an embodiment of the invention in which an object, in this example, a file, is parsed into a series of chunks for storage. Initially the file 11 is passed into the storage software where it is designated as an object 12 and allocated a unique object identification number, in this case, #007. A new entry 131 is made into the object table 13 to represent the allocation for this new object. The object is now parsed into “chunks” of data 121, 122, and 123, which are fixed-length segments of the object. Each chunk is passed through a hashing algorithm, which returns a unique hash number for the chunk. This algorithm can later

be applied to a retrieved chunk and the result compared with the original hash to ensure the retrieved chunk is the same as that stored. The hash numbers for each chunk are stored in the object table **13** in the entry row for the object **132** so that later the complete object can be retrieved by collection of the chunks.

[0058] Also in **FIG. 1**, the chunk hashes are now compared with existing entries in the chunk table **14**. Any hash that matches an existing entry **141** is already stored and so no action is taken (i.e., the data is not stored again, leading to automatic compression of the objects). A new hash (one which has no corresponding entry in the chunk table **14**) is entered into the chunk table **141**. The data in the chunk is then stored on the available storage devices **151**, **152**, and **153** in the most efficient manner for fault-tolerant storage. This approach may lead to the chunk data's being stored, for example, in a mirrored fashion on a storage system comprised of one or two devices or parity striped on a system with more than two storage devices. This data will be stored on the storage devices at physical locations **1511**, **1521**, and **1531**, and these locations and the number of locations will be stored in the chunk table columns **143** and **142** so that all physical parts of a chunk may later be located and retrieved.

[0059] **FIG. 2** illustrates in the same embodiment how a pattern for fault-tolerant storage for a chunk may be dynamically changed as a result of the addition of more storage. In particular, **FIG. 2** shows how a chunk physically stored on the storage devices may be laid out in a new pattern once additional storage is added to the overall system. In **FIG. 2(a)**, the storage system comprises two storage devices **221** and **222** and the chunk data is physically mirrored onto the two storage devices at locations **2211** and **2221** to provide fault tolerance. In **FIG. 2(b)** a third storage device **223** is added, and it becomes possible to store the chunk in a parity striped manner, a pattern which is more storage efficient than the mirrored pattern. The chunk is laid out in this new pattern in three physical locations **2311**, **2321**, and **2331**, taking a much lower proportion of the available storage. The chunk table **21** is updated to show the new layout in three locations **212** and also the new chunk physical locations **2311**, **2321**, and **2331** are recorded **213**.

[0060] **FIG. 3** shows a mature storage system, in accordance with an embodiment of the present invention, which has been functioning for some time. This illustrates how chunks may be physically stored over time on storage devices with varying storage capacities.

[0061] The figure shows a storage system comprised of a 40 GB storage device **31**, an 80 GB storage device **32** and a 120 GB storage device **33**. Initially chunks are stored in a fault tolerant stripe pattern **34** until the 40 GB storage device **31** became full. Then, due to lack of storage space, new data is stored in a mirrored pattern **36** on the available space on the 80 GB **32** and the 120 GB **33** storage devices. Once the 80 GB storage device **32** is full, then new data is laid out in a single disk fault tolerant pattern **37**. Even though the storage devices comprise a single pool for storing data, the data itself, as stored in the chunks, has been stored in a variety of distinct patterns.

[0062] **FIG. 4** illustrates another embodiment of the invention in which indicator states are used to warn of inefficient storage use and low levels of fault tolerance. In **FIG. 4(a)**, all three storage devices **41**, **42**, and **45** have free

space and the indicator light **44** is green to show data is being stored in an efficient and fault-tolerant manner. In **FIG. 4(b)** the 40 GB storage device **41** has become full, and thus new data can be stored only on the two storage devices **42** and **43** with remaining free space in a mirrored pattern **46**. In order to show the data is still fully redundant but not efficiently stored, the indicator light **44** has turned amber. In **FIG. 4(c)**, only the 120 GB storage device **43** has free space remaining and so all new data can be stored only in a mirrored pattern on this one device **43**. Because the fault-tolerance is less robust and the system is critically short of space, the indicator light **44** turns red to indicate the addition of more storage is necessary.

[0063] In one alternative embodiment, an indicator is provided for each drive/slot in the array, for example, in the form of a three-color light (e.g., green, yellow, red). In one particular embodiment, the lights are used to light the whole front of a disk carrier with a glowing effect. The lights are controlled to indicate not only the overall status of the system, but also which drive/slot requires attention (if any). Each three-color light can be placed in at least four states, specifically off, green, yellow, red. The light for a particular slot may be placed in the off state if the slot is empty and the system is operating with sufficient storage and redundancy so that no drive need be installed in the slot. The light for a particular slot may be placed in the green state if the corresponding drive is sufficient and need not be replaced. The light for a particular slot may be placed in the yellow state if system operation is degraded such that replacement of the corresponding drive with a larger drive is recommended. The light for a particular slot may be placed in the red state if the corresponding drive must be installed or replaced. Additional states could be indicated as needed or desired, for example, by flashing the light between on and off states or flashing the light between two different colors (e.g., flash between red and green after a drive has been replaced and re-layout of data is in progress). Additional details of an exemplary embodiment are described below.

[0064] Of course, other indication techniques can be used to indicate both system status and drive/slot status. For example, a single LCD display could be used to indicate system status and, if needed, a slot number that requires attention. Also, other types of indicators (e.g., a single status indicator for the system (e.g., green/yellow/red) along with either a slot indicator or a light for each slot) could be used.

[0065] **FIG. 5** is a block diagram of functional modules used in the storage, retrieval and re-layout of data in accordance with an embodiment of the invention, such as discussed above in connections with **FIGS. 1-3**. The entry and exit point for communication are the object interface **511** for passing objects to the system for storage or retrieving objects, the block interface **512**, which makes the storage system appear to be one large storage device, and the CIFS interface **513**, which makes the storage system appear to be a Windows file system. When these interfaces require the storage of data, the data is passed to the Chunk Parser **52**, which performs the break up of the data into chunks and creates an initial entry into the object table **512** (as discussed above in connection with **FIG. 1**). These chunks are then passed to the hash code generator **53**, which creates the associated hash codes for each chunk and enters these into the object table so the chunks associated with each object are listed **512** (as discussed above in connection with **FIG. 1**).

The chunk hash numbers are compared with the entries in the chunk table 531. Where a match is found, the new chunk is discarded, as it will be identical to a chunk already stored in the storage system. If the chunk is new, a new entry for it is made in the chunk table 531, and the hashed chunk is passed to the physical storage manager 54. The physical storage manager stores the chunk in the most efficient pattern possible on the available storage devices 571, 572, and 573 and makes a corresponding entry in the chunk table 531 to show where the physical storage of the chunk has occurred so that the contents of the chunk can be retrieved later 512 (as discussed above in connection with FIG. 1).

[0066] The retrieval of data in FIG. 5 by the object 511, block 512 or CIFS 513 interface is performed by a request to the retrieval manager 56, which consults the object table 521 to determine which chunks comprise the object and then requests these chunks from the physical storage manager 54. The physical storage manager 54 consults the chunk table 531 to determine where the requested chunks are stored and then retrieves them and passes the completed data (object) back to the retrieval manager 56, which returns the data to the requesting interface. Also included in FIG. 5 is the fault tolerant manager (FTL) 55, which constantly scans the chunk table to determine if chunks are stored in the most efficient manner possible. (This may change as storage devices 571, 572, and 573 are added and removed.) If a chunk is not stored in the most efficient manner possible, then the FTL will request the physical storage manager 54 to create a new layout pattern for the chunk and update the chunk table 531. This way all data continues to remain stored in the most efficient manner possible for the number of storage devices comprising the array (as discussed above in connection with FIGS. 2 and 3).

[0067] The following provides additional details of an exemplary embodiment of the present invention.

Data Layout Scheme—Zones

[0068] Among other things, a Zone has the effect of hiding redundancy and disk re-layout from the actual data being stored on the disk. Zones allow additional layout methods to be added and changed without affecting the user of the zone.

[0069] The storage array lays out data on the disk in virtual sections called Zones. A Zone stores a given and fixed amount of data (for example 1 G Bytes). A zone may reside on a single disk or span across one or more drives. The physical layout of a Zone provides redundancy in the form specified for that zone.

[0070] FIG. 6 shows an example in which mirroring is used in an array containing more than two drives. FIG. 7 shows some example zones using different layout schemes to store their data. The diagram assumes a zone stores 1 GB of data. Note the following points:

[0071] i) A zone that spans multiple drives does not necessarily use the same offset into the drive across the set.

[0072] ii) A single drive mirror requires 2 G of storage to store 1 G of data

[0073] iii) A dual drive mirror requires 2 G of storage to store 1 G of data.

[0074] iv) A 3 drive stripe requires 1.5 G of storage to store 1 G of data.

[0075] v) A 4 drive stripe requires 1.33 G of storage to store 1 G of data.

[0076] vi) Zone A, zone B etc. are arbitrary zone names. In a real implementation each zone would be identified by a unique number.

[0077] vii) Although implied by the diagram, zones are not necessarily contiguous on a disk (see regions later).

[0078] viii) There is no technical reason why mirroring is restricted to 2 drives. For example, in a 3 drive system 1 copy of the data could be stored on 1 drive and half of the mirrored data could be stored on each of the other two drives. Likewise, data could be mirrored across three drives, with half the data on each of two drives and half of the mirror on the other two drives.

Data Layout Scheme—Regions

[0079] Each disk is split into a set of equal-sized Regions. The size of a Region is much smaller than a Zone and a Zone is constructed from one or more regions from one or more disks. For efficient use of disk space, the size of a Region is typically a common factor of the different Zone sizes and the different number of disks supported by the array. In an exemplary embodiment, Regions are $\frac{1}{12}$ the data size of a Zone. The following table lists the number of Regions/Zone and the number of Regions/disk for various layouts, in accordance with an exemplary embodiment of the invention.

Layout Method	Number of regions/zone	Number of regions/disk
1 drive mirror	24	24
2 drive mirror	24	12
3 drive stripe	18	6
4 drive stripe	16	4

[0080] Individual Regions can be marked as used, free or bad. When a Zone is created, a set of free Regions from the appropriate disks are selected and logged in a table. These Regions can be in any arbitrary order and need not be contiguous on the disk. When data is written to or read from a Zone, the access is redirected to the appropriate Region. Among other things, this allows data re-layout to occur in a flexible and efficient manner. Over time, the different sized Zones will likely cause fragmentation to occur, leaving many disk areas too small to hold a complete Zone. By using the appropriate Region size, all gaps left by fragmentation will be at least one Region in size, allowing easy reuse of these small gaps with out having to de-fragment the whole disk.

Data Layout Scheme—Re-Layout

[0081] In order to facilitate implementation, a fixed sequence of expansion and contraction may be enforced. For example, if two drives are suddenly added, the expansion of a zone may go through an intermediate expansion as though one drive was added before a second expansion is performed to incorporate the second drive. Alternatively, expansion and contraction involving multiple drives may be handled atomically, without an intermediate step.

[0082] Before any re-layout occurs, the required space must be available. This should be calculated before starting the re-layout to ensure that unnecessary re-layout does not occur.

Data Layout Scheme—Drive Expansion

[0083] The following describes the general process of expanding from single drive mirroring to dual drive mirroring in accordance with an exemplary embodiment of the invention:

[0084] i) Assuming single drive mirror has data 'A' and mirror 'B'

[0085] ii) Allocate 12 regions on drive to expand zone on to 'C'

[0086] iii) Copy mirror 'B' to region set 'C'

[0087] iv) Any writes made to data already copied must be mirrored to the appropriate place in 'C'

[0088] v) When copy is complete, update zone table with new layout type and replace pointers to 'B' with pointers to 'C'

[0089] vi) Mark the regions that make-up 'B' as free.

[0090] The following describes the general process of expanding from dual drive mirroring to triple drive striping with parity in accordance with an exemplary embodiment of the invention:

[0091] i) Assume one drive has data 'A' and a second drive has mirror 'B'

[0092] ii) Allocate 6 regions on third drive for parity information 'C'

[0093] iii) Calculate parity information using first 6 regions of 'A' and the second 6 regions of 'B'

[0094] iv) Place parity information in 'C'

[0095] v) Any writes made to data already processed must be parity'd to the appropriate place in 'C'

[0096] vi) When copy is complete, update zone table with new layout type point table to first half of 'A', second half of 'B' and 'C'

[0097] vii) Mark second half of 'A' and first half of 'B' as free.

[0098] The following describes the general process of expanding from triple drive striping to quad drive striping with parity in accordance with an exemplary embodiment of the invention:

[0099] i) Assume one drive has data 'A', a second drive has data 'B' and a third has parity 'P'

[0100] ii) Allocate four regions on a fourth drive for strip data 'C'

[0101] iii) Copy last two regions of 'A' to the first two regions of 'C'

[0102] iv) Copy first two regions of 'B' to last two regions of 'C'

[0103] v) Allocate four regions on parity drive 'D'

[0104] vi) Calculate parity information using first four regions of A, C and the last four regions of B

[0105] vii) Place parity information in 'D'

[0106] viii) Any writes made to data already processed must be parity'd to the appropriate place in 'D'

[0107] ix) Update zone table with new layout type and point table to first four regions of 'A', 'C', second four regions of 'B' and 'D'

[0108] x) Mark last two regions of 'A' and first two regions of 'B' as free.

Data Layout Scheme—Drive Contraction

[0109] Drive contraction takes place when a disk is either removed or fails. In such a case, the array contracts the data to get all zones back into a redundant state if possible. Drive contraction is slightly more complex than expansion as there are more choices to make. However, transitions between layout methods happen in a similar way to expansion, but in reverse. Keeping the amount of data to be reproduced to a minimum allows redundancy to be achieved as quickly as possible. Drive contraction generally precedes one zone at a time while space is available until all zones are re-laid out. Generally speaking, only data which resides on the removed or failed disk will be rebuilt.

Choosing How to Contract

[0110] The following table describes a decision tree for each zone that needs to be re-laid out, in accordance with an exemplary embodiment of the present invention:

Zone type missing data	Condition	Action
Any	No Space available for zone re-layout	Leave zone in degraded state until new disk added or removed disk is replaced.
Single drive mirror	Data inconsistent	Lock down system and wait for reset or for the missing drive to be replaced
Dual Drive Mirror	1 disk left in system Space only available on drive that contains remaining data	Convert to single drive mirror
	2 or 3 disks left in system with space is available	Reconstruct mirror on another drive
3 Drive Striping	2 disks left in system with space available	Convert to 2 drive mirroring
	3 disks left in system with space available	Reconstruct missing stripe segment on the third drive

-continued

Zone type missing data	Condition	Action
4 Drive Striping	3 disks left in system with space available	Convert to 3 drive striping

[0111] The following describes the general process of contracting from dual drive mirroring to single drive mirroring in accordance with an exemplary embodiment of the invention:

[0112] i) Assuming single drive mirror has data 'A' and missing mirror 'B' or visa versa

[0113] ii) Allocate 12 regions on the drive that contains 'A' as 'C'

[0114] iii) Copy data 'A' to region set 'C'

[0115] iv) Any writes made to data already copied must be mirrored to the appropriate place in 'C'

[0116] v) When copy is complete, update zone table with new layout type and replace pointers to 'B' with pointers to 'C'

[0117] The following describes the general process of contracting from triple drive stripe to dual drive mirror (missing parity) in accordance with an exemplary embodiment of the invention:

[0118] i) Assuming that stripe consists of data blocks 'A', 'B' and 'C' on different drives. Parity 'C' is missing.

[0119] ii) Define 'A' as containing the first half of the zone and 'B' as the second half of the zone.

[0120] iii) Allocate 6 regions 'D' on the 'A' drive and 6 regions 'E' on the 'B' drive

[0121] iv) Copy 'A' to 'E'.

[0122] v) Copy 'B' to 'D'

[0123] vi) Any writes made to data already copied must be mirrored to the appropriate place in 'D' and 'E'

[0124] vii) When copy is complete, update zone table with new layout type and set pointers to 'A'/'D' and 'E'/'B'

[0125] The following describes the general process of contracting from triple drive stripe to dual drive mirror (missing data) in accordance with an exemplary embodiment of the invention:

[0126] i) Assuming that stripe consists of data blocks 'A', 'B' and 'C' on different drives. Data 'C' is missing.

[0127] ii) Define 'A' as containing the first half of the zone and 'C' as the second half of the zone.

[0128] iii) Allocate 6 regions 'D' on the 'A' drive and 12 regions 'E' on the 'B' drive

[0129] iv) Copy 'A' to the first half of 'E'

[0130] v) Reconstruct missing data from 'A' and 'B'. Write data to 'D'

[0131] vi) Copy 'D' to second half of 'E'.

[0132] vii) Any writes made to data already copied must be mirrored to the appropriate place in 'D' and 'E'

[0133] viii) When copy is complete, update zone table with new layout type and set pointers to 'A'/'D' and 'E'

[0134] ix) Mark 'B' regions as free.

[0135] The following describes the general process of contracting from quad drive stripe to triple drive stripe (missing parity) in accordance with an exemplary embodiment of the invention:

[0136] i) Assuming that stripe consists of data blocks 'A', 'B', 'C' and 'D' on different drives. Parity 'D' is missing.

[0137] ii) Define 'A' as containing the first third, 'B' as the second third and 'C' as the third third of the zone

[0138] iii) Allocate 2 regions 'G' on the 'A' drive, 2 regions 'E' on the 'C' drive and 6 regions 'F' on the 'B' drive.

[0139] iv) Copy first half of 'B' to 'G'

[0140] v) Copy second half of 'B' to 'E'

[0141] vi) Construct parity from 'A'/'G' and 'E'/'C' and write to 'F'

[0142] vii) Any writes made to data already copied must be mirrored to the appropriate place in 'G', 'E' and 'F'

[0143] viii) When copy is complete, update zone table with new layout type and set pointers to 'A'/'G', 'E'/'C' and 'F'

[0144] ix) Mark 'B' regions as free.

[0145] The following describes the general process of contracting from quad drive stripe to triple drive stripe (first 1/3 missing) in accordance with an exemplary embodiment of the invention:

[0146] i) Assuming that stripe consists of data blocks 'A', 'B', 'C' and 'D' on different drives. Data 'A' is missing.

[0147] ii) Define 'A' as containing the 1st third, 'B' as the 2nd third and 'C' as the 3rd third of the zone and 'D' as the parity.

[0148] iii) Allocate 4 regions 'E' on the 'B' drive, 2 regions 'F' on the 'C' drive and 6 regions 'G' on the 'D' drive.

[0149] iv) Copy second half of 'B' to 'F'

[0150] v) Construct missing data from 'B', 'C' and 'D' and write to 'E'

[0151] vi) Construct new parity from 'E'/'1st half B' and 'F'/'C' and write to 'G'

[0152] vii) Any writes made to data already copied must be mirrored to the appropriate place in 'B', 'E', 'F' and 'G'

[0153] viii) When copy is complete, update zone table with new layout type and set pointers to 'E'/1st half 'B' and 'F'/'C' and 'G'

[0154] ix) Mark 2nd half 'B' and 'D' regions as free.

[0155] The following describes the general process of contracting from quad drive stripe to triple drive stripe (second $\frac{1}{3}$ missing) in accordance with an exemplary embodiment of the invention:

[0156] i) Assuming that stripe consists of data blocks 'A', 'B', 'C' and 'D' on different drives. Data 'B' is missing.

[0157] ii) Define 'A' as containing the 1st third, 'B' as the 2nd third and 'C' as the 3rd third of the zone and 'D' as the parity.

[0158] iii) Allocate 2 regions 'E' on the 'A' drive, 2 regions 'F' on the 'C' drive and 6 regions 'G' on the 'D' drive.

[0159] iv) Construct missing data from 1st half 'A', 1st half 'C' and 1st half 'D' and write to 'E'

[0160] v) Construct missing data from 2nd half 'A', 2nd half 'C' and 2nd half 'D' and write to 'F'

[0161] vi) Construct new parity from 'A'/'E' and 'F'/'C' and write to 'G'

[0162] vii) Any writes made to data already copied must be mirrored to the appropriate place in 'E', 'F' and 'G'

[0163] viii) When copy is complete, update zone table with new layout type and set pointers to 'E', 'F' and 'G'

[0164] ix) Mark 'D' regions as free.

[0165] The following describes the general process of contracting from quad drive stripe to triple drive stripe (third $\frac{1}{3}$ missing) in accordance with an exemplary embodiment of the invention:

[0166] i) Assuming that stripe consists of data blocks 'A', 'B', 'C' and 'D' on different drives. Data 'C' is missing.

[0167] ii) Define 'A' as containing the 1st third, 'B' as the 2nd third and 'C' as the 3rd third of the zone and 'D' as the parity.

[0168] iii) Allocate 2 regions 'E' on the 'A' drive, 4 regions 'F' on the 'B' drive and 6 regions 'G' on the 'D' drive.

[0169] iv) Copy 1st half 'B' to 'E'

[0170] v) Construct missing data from 'A', 'B' and 'D' and write to 'F'

[0171] vi) Construct new parity from 'A'/'E' and 2nd half 'B'/'F' and write to 'G'

[0172] vii) Any writes made to data already copied must be mirrored to the appropriate place in 'E', 'F' and 'G'

[0173] viii) When copy is complete, update zone table with new layout type and set pointers to 'A'/'E' and 2nd half 'B'/'F' and 'G'

[0174] ix) Mark 1st half 'B' and 'D' regions as free.

[0175] For example, with reference again to FIG. 3, dual drive mirror (Zone B) could be reconstructed on Drive 2 if either Drive 0 or Drive 1 is lost, provided there is sufficient space available on Drive 2. Similarly, three drive stripe

(Zone C) could be reconstructed utilizing Drive 3 if any of Drives 0-2 are lost, provided there is sufficient space available on Drive 3.

Data Layout Scheme—Zone Reconstruction

[0176] Zone reconstruction occurs when a drive has been removed and there is enough space on the remaining drives for ideal zone re-layout or the drive has been replaced with a new drive of larger size.

[0177] The following describes the general process of dual drive mirror reconstruction in accordance with an exemplary embodiment of the invention:

[0178] i) Assuming single drive mirror has data 'A' and missing mirror 'B'

[0179] ii) Allocate 12 regions 'C' on a drive other than that containing 'A'

[0180] iii) Copy data 'A' to 'C'

[0181] iv) Any writes made to data already copied must be mirrored to the appropriate place in 'C'

[0182] v) When copy is complete, update zone table pointers to 'B' with pointers to 'C'

[0183] The following describes the general process of three drive stripe reconstruction in accordance with an exemplary embodiment of the invention:

[0184] i) Assume one drive has data 'A', a second drive has data 'B' and a third has parity 'P'. 'B' is missing. Note it doesn't matter which piece is missing, the required action is the same in all cases.

[0185] ii) Allocate 6 regions 'D' on a drive other than that containing 'A' and 'P'

[0186] iii) Construct missing data from 'A' and 'P'. Write data to 'D'

[0187] iv) Any writes made to data already processed must be parity'd to the appropriate place in 'D'

[0188] v) Update zone table by replacing pointers to 'B' with pointers to 'D'

[0189] In this exemplary embodiment, four -drive-reconstruction can only occur if the removed drive is replaced by another drive. The reconstruction consists of allocating six regions on the new drive and reconstructing the missing data from the other three region sets.

Data Layout Scheme—The Temporarily Missing Drive Problem

[0190] When a drive is removed and there is no room for re-layout, the array will continue to operate in degraded mode until either the old drive is plugged back in or the drive is replaced with a new one. If a new one is plugged in, then the drive set should be rebuilt. In this case, data will be re-laid out. If the old disk is placed back into the array, it will no longer be part of the current disk set and will be regarded as a new disk. However, if a new disk is not placed in the array and the old one is put back in, the old one will still be recognized as being a member of the disk set, albeit an out of date member. In this case, any zones that have already been re-laid out will keep their new configuration and the regions on the old disk will be freed. Any zone that has not been re-laid out will still be pointing at the appropriate

regions of the old disk. However, as some writes may have been performed to the degraded zones, these zones need to be refreshed. Rather than logging every write that has occurred, degraded regions that have been modified may be marked. In this way, when the disk is replaced, only the regions that have been modified need to be refreshed.

[0191] Furthermore, zones that have been written to may be placed further up the priority list for re-layout. This should reduce the number of regions that need to be refreshed should the disk be replaced. A timeout may also be used, after which point the disk, even if replaced, will be wiped. However, this timeout could be quite large, possibly hours rather than minutes.

Data Layout Scheme—Data Integrity

[0192] As discussed above, one problem with a standard RAID system is that it is possible for disc-surface corruption to occur on an infrequently used area of the disk array. In the event that another drive fails, it is not always possible to determine that corruption has occurred. In such a case, the corrupted data may be propagated and preserved when the RAID array rebuilds the failed drive.

[0193] The hash mechanism discussed above provides an additional mechanism for data corruption detection over that which is available under RAID. As is mentioned elsewhere, when a chunk is stored, a hash value is computed for the chunk and stored. Any time the chunk is read, a hash value for the retrieved chunk can be computed and compared with the stored hash value. If the hash values do not match (indicating a corrupted chunk), then chunk data can be recovered from redundant data.

[0194] In order to minimize the time window in which data corruption on the disk can occur, a regular scan of the disks will be performed to find and correct corrupted data as soon as possible. It will also, optionally, allow a check to be performed on reads from the array.

Data Layout Scheme—Volume

[0195] In a sparse volume, regardless of the amount of storage space available on discs in the array, the array always claims to be a fixed size—for example, M Terabytes. Assume that the array contains S bytes of actual storage space, where $S \leq M$, and that data can be requested to be stored at locations L1, L2, L3, etc. within the M Terabyte space. If the requested location $L_n > S$, then the data for L_n must be stored at a location $P_n < S$. This is managed by including a lookup table to index P_n based on L_n , as shown in FIG. 8. The feature allows the array to work with operating systems that do not support volume expansion, such as Windows, Linux, and Apple operating systems. In addition, the array can provide multiple Sparse Volumes which all share the same physical storage. Each sparse volume will have a dedicated lookup table, but will share the same physical space for data storage.

Drive Slot Indicators

[0196] As discussed above, the storage array consists of one or more drive slots. Each drive slot can either be empty or contain a hard disk drive. Each drive slot has a dedicated indicator capable of indicating four states: Off, OK, Degraded and Fail. The states are interpreted generally as follows:

Indicator State	Meaning for Array User
Off	Drive slot is empty and is available for an additional drive to be inserted.
OK	Drive in slot is functioning correctly.
Degraded	Action by user recommend: if slot is empty, add a drive to this slot; if slot contains a drive, replace drive with another, higher-capacity drive.
Fail	Action by user required ASAP: if slot is empty, add a drive to this slot; if slot contains a drive, replace drive with another, higher-capacity drive.

[0197] In this exemplary embodiment, red/amber/green light emitting diodes (LEDs) are used as the indicators. The LEDs are interpreted generally as follows:

LED state	Indicator State	Example circumstances under which state may occur	FIGS.
Off	Off	Slot is empty. Array has available space.	9, 10, 12
Green	OK	Drive is functioning correctly, array data is redundant and array has available disk space.	9, 10, 11, 12
Amber	Degraded	Array is approaching a Fail condition; Not enough space to maintain redundant data in the event of a disc failure.	11
Red	Fail	Disk in this slot has failed and must be replaced; the array does not have enough space to maintain redundant data storage and more space must be added.	10, 12

[0198] FIG. 9 shows an exemplary array having available storage space and operating in a fault-tolerant manner, in accordance with an exemplary embodiment of the present invention. Slots B, C, and D are populated with storage devices, and there is sufficient storage space available to store additional data redundantly. The indicators for slots B, C, and D are green (indicating that these storage devices are operating correctly, the array data is redundant, and the array has available disk space), and the indicator for slot A is off (indicating that no storage device needs to be populated in slot A).

[0199] FIG. 10 shows an exemplary array that does not have enough space to maintain redundant data storage and more space must be added, in accordance with an exemplary embodiment of the present invention. Slots B, C, and D are populated with storage devices. The storage devices in slots C and D are full. The indicators for slots B, C, and D are green (indicating that these storage devices are operating correctly), and the indicator for slot A is red (indicating that the array does not have enough space to maintain redundant data storage and a storage device should be populated in slot A).

[0200] FIG. 11 shows an exemplary array that would be unable to maintain redundant data in the event of a failure, in accordance with an exemplary embodiment of the present invention. Slots A, B, C, and D are populated with storage devices. The storage devices in slots C and D are full. The

indicators for slots A, B, and C are green (indicating that they are operating correctly), and the indicator for slot D is amber (indicating that the storage device in slot D should be replaced with a storage device having greater storage capacity).

[0201] FIG. 12 shows an exemplary array in which a storage device has failed, in accordance with an exemplary embodiment of the present invention. Slots B, C, and D are populated with storage devices. The storage device in slot C has failed. The indicators for slots B and D are green (indicating that they are operating correctly), the indicator for slot C is red (indicating that the storage device in slot C should be replaced), and the indicator for slot A is off (indicating that no storage device needs to be populated in slot A).

[0202] The following is a description of the software design for an exemplary embodiment of the present invention. The software design is based on six software layers, which span the logical architecture from physically accessing the disks to communicating with the host computing system.

[0203] In this exemplary embodiment, a file system resides on a host server, such as a Windows, Linux, or Apple server, and accesses the storage array as a USB or iSCSI device. Physical disk requests arriving over the host interface are processed by the Host Request Manager (HRM). A Host I/O interface coordinates the presentation of a host USB or iSCSI interface to the host, and interfaces with the HRM. The HRM coordinates data read/write requests from the host I/O interface, dispatches read and write requests, and co-ordinates the retiring of these requests back to the host as they are completed.

[0204] An overarching aim of the storage array is to ensure that once data is accepted by the system, it is stored in a reliable fashion, making use of the maximum amount of redundancy the system currently stores. As the array changes physical configuration, data is re-organized so as to maintain (and possibly maximize) redundancy. In addition, simple hash based compression is used to reduce the amount of storage used.

[0205] The most basic layer consists of disk drivers to store data on different disks. Disks may be attached via various interfaces, such as ATA tunneled over a USB interface.

[0206] Sectors on the disks are organized into regions, zones, and clusters, each of which has a different logical role.

[0207] Regions represent a set of contiguous physical blocks on a disk. On a four drive system, each region is 1/2 GB in size, and represents minimal unit of redundancy. If a sector in a region is found to be physically damaged, the whole region will be abandoned.

[0208] Zones represent units of redundancy. A zone will consist of a number of regions, possibly on different disks to provide the appropriate amount of redundancy. Zones will provide 1 GB of data capacity, but may require more regions in order to provide the redundancy. 1 GB with no redundancy require one set of 12 regions (1 GB); a 1 GB mirrored zone will require 2 sets of 1 GB regions (24 regions); a 1 GB

3-disk stripped zone will require 3 sets of 0.5 GB regions (18 regions). Different zones will have different redundancy characteristics.

[0209] Clusters represent the basic unit of compression, and are the unit size within zones. They are currently 4 KB: 8x512 byte sectors in size. Many clusters on a disk will likely contain the same data. A cluster access table (CAT) is used to track the usage of clusters via a hashing function. The CAT translates between logical host address and the location of the appropriate cluster in the zone.

[0210] When writing to disk, a hash function is used to see if the data is already present on the disk. If so, the appropriate entry in the CAT table is set to point to the existing cluster.

[0211] The CAT table resides in its own zone. If it exceeds the size of the zone, an additional zone will be used, and a table will be used to map logical address to the zone for that part of the CAT. Alternatively, zones are pre-allocated to contain the CAT table.

[0212] In order to reduce host write latency and to ensure data reliability, a journal manager will record all write requests (either to disk, or to NVRAM). If the system is rebooted, journal entries will be committed on reboot.

[0213] Disks may come and go, or regions may be retired if they are found to have corruption. In either of these situations, a layout manager will be able to re-organize regions within a zone in order to change its redundancy type, or change the regional composition of a zone (should a region be corrupted).

[0214] As the storage array provides a virtual disk array, backed by changing levels of physical disk space, and because it presents a block level interface, it is not obvious when clusters are no longer in use by the file system. As a result, the cluster space used will continue to expand. A garbage collector (either located on the host or in firmware) will analyze the file system to determine which clusters have been freed, and remove them from the hash table.

[0215] The following table shows the six software layers in accordance with this exemplary embodiment of the invention:

Layer 5:	Garbage collector, Host Interface (USB/iSCSI)
Layer 4:	Host request manager
Layer 3:	CAT, HASH, Journal manager
Layer 2:	Zones manager. Allocates/frees chunks of sectors called Zones. Knows about SDM, DDM, SD3 etc in order to deal with errors and error recovery. Layout Manager
Layer 1:	Read/write physical clusters/sectors. Allocates Regions per disk
Layer 0:	Disk access drivers

[0216] FIG. 13 shows a module hierarchy representing the different software layers and how they relate to one another. Software layering is preferably rigid in order to present clear APIs and delineation.

[0217] The Garbage Collector frees up clusters which are no longer used by the host file system. For example, when a file is deleted, the clusters that were used to contain the file are preferably freed.

[0218] The Journal Manager provides a form of journaling of writes so that pending writes are not lost in the case of a power failure or other error condition.

[0219] The Layout Manager provides run-time re-layout of the Zones vis-à-vis their Regions. This may occur as a result of disk insertion/removal or failure.

[0220] The Cluster Manager allocates clusters within the set of data Zones. The Disk Utilization Daemon checks for free disk space on a periodic basis.

[0221] The Lock Table deals with read after write collision issues.

[0222] The Host Request Manager deals with the read/write requests from the Host and Garbage Collector. Writes are passed to the Journal Manager, whereas Reads are processed via the Cluster Access Table (CAT) Management layer.

[0223] As discussed above, in typical file systems, some amount of the data will generally be repetitive in nature. In order to reduce disk space utilization, multiple copies of this data are not written out to the disks. Instead, one instance is written, and all other instances of the same data are referenced to this one instance.

[0224] In this exemplary embodiment, the system operates on a cluster of data at any time (e.g., 8 physical sectors), and this is the unit that is hashed. The SHA1 algorithm is used to generate a 160-bit hash. This has a number of benefits, including good uniqueness, and being supported on-chip in a number of processors. All 160-bits will be stored in the hash record, but only the least significant 16-bits will be used as an index into a hash table. Other instances matching the lowest 16-bits will be chained via a linked-list.

[0225] In this exemplary embodiment, only one read/write operation may occur at a time. For performance purposes, hash analysis is not permitted to happen when writing a cluster to disk. Instead, hash analysis will occur as a background activity by the hash manager.

[0226] Write requests are read from the journal's write queue, and are processed to completion. In order to ensure data consistency, writes must be delayed if a write operation is already active on the cluster. Operations on other clusters may proceed unimpeded.

[0227] Unless a whole cluster is being written, the data being written will need to be merged with the existing data stored in the cluster. Based on the logical sector address (LSA), the CAT entry for the cluster is located. The hash key, zone and cluster offset information is obtained from this record, which can then be used to search the hash table to find a match. This is the cluster.

[0228] It might well be necessary to doubly hash the hash table; once via the SHA1 digest, and then by the zone/cluster offset to improve the speed of lookup of the correct hash entry. If the hash record has already been used, the reference count is decremented. If the reference count is now zero, and there is no snapshot referenced by the hash entry, the hash entry and cluster can be freed back to their respective free lists.

[0229] The original cluster data is now merged with the update section of the cluster, and the data is re-hashed. A new cluster is taken off the free-list, the merged data is

written to the cluster, new entry is added to the hash table, and the entry in the CAT table is updated to point to the new cluster.

[0230] As a result of updating the hash table, the entry is also added to an internal queue to be processed by a background task. This task will compare the newly added cluster and hash entry with other hash entries that match the hash table row address, and will combine records if they are duplicates, freeing up hash entries and CAT table entries as appropriate. This ensures that write latency is not burdened by this activity. If a failure (e.g., a loss of power) occurs during this processing, the various tables can be deleted, with a resulting loss of data. The tables should be managed in such a way that the final commit is atomic or the journal entry can be re-run if it did not complete fully.

[0231] The following is pseudocode for the write logic:

```

While (stuff to do)
writeRecord = journalMgr.read( );
lsa = writeRecord.GetLsa( );
catEntry = catMgr.GetCATEntry(lsa);
if (catMgr.writeInProgress(catEntry)) delay( );
originalCluster = catMgr.readCluster(catEntry);
originalHash = hashMgr.calcHash(originalCluster);
hashRecord = hashMgr.Lookup(originalHash, zone, offset);
if ((hashRecord.RefCount == 1) && (hashRecord.snapshot == 0))
    hashRecord.free( );
    originalCluster.free( );
    // Note there are some optimizations here where we can reuse
    // this cluster without having to free & re-allocate it.
// otherwise, still users of this cluster, so update & leave it alone
hashRecord.RefCount--;
hashRecord.Update(hashRecord);
// Now add new record
mergedCluster = mergeCluster(originalCluster, newCluster);
newHash = hashMgr.calcHash(mergedCluster);
newCluster = clusterMgr.AllocateCluster(zone, offset);
clusterMgr.write(cluster, mergedCluster);
zoneMgr.write(cluster, mergedCluster);
...
hashMgr.addHash(newHash, newCluster, zone, offset)
    (internal: queue new hash for background processing)
catMgr.Update(lsa, zone, offset, newHash);
// We've completed the journal entry successfully. Move on to the next
one.
journalMgr.next( );

```

[0232] Read requests are also processed one cluster (as opposed to "sector") at a time. Read requests do not go through the hash-related processing outlined above. Instead, the host logical sector address is used to reference the CAT and obtain a Zone number and cluster offset into the Zone. Read requests should look up the CAT table entry in the CAT Cache, and must be delayed in the write-in-progress bit is set. Other reads/writes may proceed unimpeded. In order to improve data integrity checking, when a cluster is read, it will be hashed, and the hash compared with the SHA1 hash value stored in the hash record. This will require using the hash, zone and cluster offset as a search key into the hash table.

[0233] Clusters are allocated to use as few Zones as possible. This is because Zones correspond directly to disk drive usage. For every Zone, there are two or more Regions on the hard drive array. By minimizing the number of Zones, the number of physical Regions is minimized and hence the consumption of space on the hard drive array is reduced.

[0234] The Cluster Manager allocates cluster from the set of Data Zones. A linked list is used to keep track of free clusters in a Zone. However, the free cluster information is stored as a bit map (32 KB per Zone) on disk. The linked list is constructed dynamically from the bitmap. Initially, a linked list of a certain number of free clusters is created in memory. When clusters are allocated, the list shrinks. At a predetermined low-water mark, new linked list nodes representing free clusters are extracted from the bitmap on disk. In this way, the bitmap does not need to be parsed in order to find a free cluster for allocation.

[0235] In this exemplary embodiment, the hash table is a 64K table of records (indexed by the lower 16 bits of the hash) and has the following format:

Offset	Size in bits	Name	Value/Valid Range	Description
0	160	sha1Hash		The complete SHA1 hash digest
	16	refCount		Number of instances of this hash; what do we do if we get beyond 16 bits?
	18	Cluster offset		Cluster offset within zone
	14	Zone #		Zone# containing this cluster
	8	snapshot		One bit per snapshot instance to indicate that this cluster entry is used by that snapshot. This model supports 8 snapshots (possible only 7)

[0236] A cluster of all zeros may be fairly common, so the all-zeros case may be treated as a special case, for example, such that it can never be deleted (so wrapping the count would not be a problem).

[0237] A linked list of free hash record is used when the multiple hashes have the same least significant hash, or when two hash entries point to different data clusters. In either case, a free hash record will be taken from the list, and linked via the pNextHash pointer.

[0238] The hash manager will tidy up entries added to the hash table, and will combine identical clusters on the disk. As new hash records are added to the hash table, a message will be posted to the hash manager. This will be done automatically by the hash manager. As a background activity, the hash manager will process entries on its queue. It will compare the full hash value to see if it matches any existing hash records. If it does, it will also compare the complete cluster data. If the clusters match, the new hash record can be discarded back to the free queue, the hash record count will be incremented, and the duplicate cluster will be returned to the cluster free queue. The hash manager must take care to propagate the snapshot bit forward when combining records.

[0239] A Cluster Access Table (CAT) contains indirect pointers. The pointers point to data clusters (with 0 being the first data cluster) within Zones. One CAT entry references a single data cluster (tentatively 4 KB in size). CATs are used (in conjunction with hashing) in order to reduce the disk usage requirements when there is a lot of repetitive data. A single CAT always represents a contiguous block of storage. CATs are contained within non-data Zones. Each CAT entry

is 48-bits. The following table shows how each entry is laid out (assuming each data Zone contains 1 GB of data):

Bits 0–17	Bits 18–31	Bits 32–47	Bits 48–63[...]
Offset of data cluster within Zone	Zone# containing data	Hash key	Reserved. Candidates include garbage collector write-bit; snapshot bits; snapshot table hash key

[0240] It is desirable for the CAT to fit into 64 bits, but this is not a necessity. The CAT table for a 2TB array is currently ~4 GB in size. Each CAT entry points to a Zone which contains the data and the number of the Zone.

[0241] FIG. 14 shows how the CAT is used to access a data clusters in a Zone. Redundant data is referenced by more than one entry in the CAT. Two logical clusters contain the same data, so their CAT entries are pointed to the same physical cluster.

[0242] The Hash Key entry contains the 16-bit extract of the 160-bit SHA1 hash value of the entire cluster. This entry is used to update the hash table during a write operation.

[0243] There are enough bits in each entry in the CAT to reference 16TB of data. However, if every data cluster is different from another (in terms of contents), then just over 3 Zones' worth of CAT entries are required to reference 2 TB of data (each zone is 1 GB in size, and hence can store 1 GB/size of CAT entry entries. Assuming 6 byte CAT entries, that is 178956970 CAT entries/zone, i.e. the table references around 682 GB/zone if each cluster is 4K).

[0244] A Host Logical Sector Translation Table is used to translate a Host Logical Sector Address into a Zone number. The portion of the CAT that corresponds to the Host Logical Sector Address will reside in this zone. Note that each CAT entry represents a cluster size of 4096 bytes. This is eight 512 byte sectors. The following shows a representation of the host logical sector translation table:

Start Host Logical Sector Address	End Host Logical Sector Address	Zone # of CAT
0 (cluster #0)	1431655759 (cluster #178956969)	
1431655760 (cluster #178956970)	...	

[0245] Zones can be pre-allocated to hold the entire CAT. Alternatively, Zones can be allocated for the CAT as more entries to the CAT are required. Since the CAT maps the 2TB virtual disk to the host sector address space, it is likely that a large part of the CAT will be referenced during hard disk partitioning or formatting by the host. Because of this, the Zones may be pre-allocated.

[0246] The CAT is a large 1 GB/zone table. The working set of clusters being used will be a sparse set from this large table. For performance reasons, active entries (probably temporally) may be cached in processor memory rather than

always reading them from the disk. There are at least two options for populating the cache—individual entries from the CAT, or whole clusters from the CAT.

[0247] Because the write-in-progress is combined with the CAT cache table, it is necessary to ensure that all outstanding writes remain in the cache. Therefore, the cache needs to be at least as large at the maximum number of outstanding write requests.

[0248] Entries in the cache will be a cluster size (ie. 4K). There is a need to know whether there is a write-in-progress in operation on a cluster. This indication can be stored as a flag in the cache entry for the cluster. The following table shows the format of a CAT cache entry:

Bits 0–17	Bits 18–31	Bits 32–47	Bit 48–63
Offset of data cluster within Zone	Zone# containing data	Hash key	Bit 48: Write-in-progress Bit 49: Dirty

[0249] The write-in-progress flag in the cache entry has two implications. First, it indicates that a write is in progress, and any reads (or additional writes) on this cluster must be held off until the write has completed. Secondly, this entry in the cache must not be flushed while the bit is set. This is partly to protect the state of the bit, and also to reflect the fact that this cluster is currently in use. In addition, this means that the size of the cache must be at least as large as the number of outstanding write operations.

[0250] One advantage of storing the write-in-progress indicator in the cache entry for the cluster is that it reflects the fact that the operation is current, saves having another table, and it saves an additional hashed-based lookup, or table walk to check this bit too. The cache can be a write-delayed cache. It is only necessary to write a cache entry back to disk when the write operation has completed, although it might be beneficial to have it written back earlier. A hash function or other mechanism could be used to increase the number of outstanding write entries that can be hashed.

[0251] An alternate approach is to cache whole clusters of the CAT (i.e., 4K entry of entries). This would generally help performance if there is good spatial locality of access. Care needs to be taken because CAT entries are 48 bits wide, so there will not be a whole number of entries in the cache. The following table shows an example of a clustered CAT cache entry:

2 words	2 words	2 words	2 words
CAT entry 1 (partial entry of last 2 words)		CAT Entry 2	
	CAT Entry 3		CAT Entry 4
	CAT entry 4		CAT entry 5
	CAT Entry 5		CAT Entry 6
...			
	CAT Entry 682		CAT Entry 683 (partial entry of first 2 words)

-continued

2 words	2 words	2 words	2 words
Write-in-progress bit array [682 bits]; bits 0–255			
Write-in-progress bit array bits 256–511			
Write-in-progress bit array 512–682 + spare bits	Dirty count	Reserved	

[0252] The table size would be 4096+96 (4192 bytes). Assuming it is necessary to have a cache size of 250 entries, the cache would occupy approximately 1 MB.

[0253] It is possible to calculate whether the first and last entry is incomplete or not by appropriate masking of the logical CAT entry address. The caching lookup routine should do this prior to loading an entry and should load the required CAT cluster.

[0254] When the host sends a sector (or cluster) read request, it sends over the logical sector address. The logical sector address is used as an offset into the CAT in order to obtain the offset of the cluster in the Zone that contains the actual data that is requested by the host. The result is a Zone number and an offset into that Zone. That information is passed to the Layer 2 software, which then extracts the raw cluster(s) from the drive(s).

[0255] In order to deal with clusters that have never been written to by the host, all CAT entries are initialized to point to a “Default” cluster which contain all zeros.

[0256] The journal manager is a bi-level write journaling system. An aim of the system is to ensure that write requests can be accepted from the host and quickly indicate back to the host that the data has been accepted while ensuring its integrity. In addition, the system needs to ensure that there will be no corruption or loss of any block level data or system metadata (e.g., CAT and Hash table entries) in the event of a system reset during any disk write.

[0257] The J1 journal manager caches all write requests from the hosts to disk as quickly as possible. Once the write has successfully completed (i.e., the data has been accepted by the array), the host can be signaled to indicate that the operation has completed. The journal entry allows recovery of write requests when recovering from a failure. Journal records consist of the data to be written to disk, and the meta-data associated with the write transaction.

[0258] In order to reduce disk read/writes, the data associated with the write will be written to free clusters. This will automatically mirror the data. Free clusters will be taken from the free cluster list. Once the data is written, the free cluster list must be written back to disk.

[0259] A journal record will be written to a journal queue on a non-mirrored zone. Each record will be a sector in size, and aligned to a sector boundary in order to reduce the risk that a failure during a journal write would corrupt a previous journal entry. Journal entries will contain a unique, incrementing sequence count at the end of the record so that the end of a queue can easily be identified.

[0260] Journal write operations will happen synchronously within a host queue processing thread. Journal writes must be ordered as they are written to disk, so only one thread may write to the journal as any time. The address of the journal entry in the J1 table can be used as a unique identifier so that the J1 journal entry can be correlated with

entries in the J2 journal. Once the journal entry is written, a transaction completion notification will be posted to the host completion queue. Now the write operation can be executed. It is important to ensure that any subsequent reads to a cluster before the journal write has completed are delayed.

[0261] The following table shows the format of the J2 journal record:

Size in bits	Name	Details
32	LBA	Logical Block Address
14	Zone	Zone # of associated cluster
18	Offset	Cluster offset of associated cluster
16	Size	Size of data
16	SequenceNumber	An incrementing sequence number so we can easily find the end of the queue

[0262] Each journal record will be aligned to a sector boundary. A journal record might contain an array of zone/offset/size tuples.

[0263] FIG. 15 shows a journal table update in accordance with an exemplary embodiment of the present invention. Specifically, when a host write request is received, the journal table is updated, one or more clusters are allocated, and data is written to the cluster(s).

[0264] Host journal requests are processed. These cause clusters to be written, and also cause updates to meta-data structure which must be shadowed back to disk (for example, the CAT table). It is important to ensure that these meta-data structures are correctly written back to disk even if a system reset occurs. A low level disk I/O write (J2) journal will be used for this.

[0265] In order to process a host interface journal entry, the appropriate manipulation of meta-data structures should be determined. The changes should occur in memory and a record of changes to various disk blocks be generated. This record contains the actual changes on disk that should be made. Each data structure that is updated is registered with the J2 journal manager. This record should be recorded to a disk based journal, and stamped with an identifier. Where the record is connected with a J1 journal entry, the identifiers should be linked. Once the record is stored, the changes to disk can be made (or can be done via a background task).

[0266] The J2 journal exists logically at layer 3. It is used to journal meta-data updates that would involve writes through the zone manager. When playback of a journal entry occurs, it will use zone manager methods. The journal itself can be stored in a specialized region. Given the short lifespan of journal entries, they will not be mirrored. Not all meta-data updates need to go through the J2 journal, particularly if updates to structures are atomic. The region manager structure may not use the J2 journal. It would be possible to detect inconsistencies in the region manager bitmap, for example, with an integrity checking background thread.

[0267] A simple approach for the J2 journal is to contain a single record. As soon as the record is committed to disk, it is replayed, updating the structures on disk. It is possible to have multiple J2 records and to have a background task committing updating records on disks. In this case, close

attention will need to be paid to the interaction between the journal and any caching algorithms associated with the various data structures.

[0268] The initial approach will run the journal entry as soon as it has been committed to disk. In principle there could be multiple concurrent users of the J2, but the J2 journal may be locked to one user at a time. Even in this case, journal entries should be committed as soon as they have been submitted.

[0269] It is important to ensure that the meta-data structures are repaired before any higher level journal activity occurs. On system reboot, the J2 journal is analyzed, and any records will be replayed. If a journal entry is correlated with a J1 journal entry, the J1 entry will be marked as completed, and can be removed. Once all J2 journal entries have been completed, the meta-data is in a reliable state and any remaining J1 journal entries can be processed.

[0270] The J2 journal record includes the following information:

[0271] Number of operations

[0272] Each operation contains:

[0273] J1 record indicator

[0274] Zone/Data offset to write to

[0275] Data to write

[0276] Size of data

[0277] Offset into data cluster

[0278] Journal record identifier

[0279] End Marker

[0280] This scheme could operate similarly to the J1 journal scheme, for example, with a sequence number to identify the end of a J2 journal entry and placing J2 journal entries on sector boundaries.

[0281] If the J1 data pointer indicator is set, then this specific operation would point to a J1 journal record. The host supplied write data would not have to be copied into our journal entry. The operation array should be able to be defined as a fixed size as the maximum number of operations in a journal record is expected to be well understood.

[0282] In order to permit recovery from corruption of a sector during a low level write operation (e.g., due to a loss of power), the J2 journal could store the whole sector that was written so that the sector could be re-written from this information if necessary. Alternatively or additionally, a CRC calculated for each modified sector could be stored in the J2 record and compared with a CRC computed from the sector on disk (e.g., by the zone manager) in order to determine whether a replay of the write operation is required.

[0283] The different journals can be stored in different locations, so there will be an interface layer provided to write journal records to backing store. The location should be non-volatile. Two candidates are hard disk and NVRAM. If the J1 journal is stored to hard disk, it will be stored in a J1 journal non-mirrored zone. The J1 journal is a candidate for storing in NVRAM. The J2 journal should be stored on disk, although it can be stored in a specialized region (i.e., not redundant, as it has a short lifespan). An advantage of storing the J2 journal on disk is that, if there is a system reset during an internal data structure update, the data structures

can be returned to a consistent state (even if the unit is left un-powered for a long period of time).

[0284] The Zones Manager (ZM) allocates Zones that are needed by higher level software. Requests to the ZM include:

[0285] a. Allocate Zone

[0286] b. De-allocate/Free Zone

[0287] c. Control data read/write pass through to L1 (?)

[0288] d. Read/Write cluster in a Zone (given the offset of the cluster and the Zone number)

[0289] The ZM manages the redundancy mechanisms (as a function of the number of drives and their relative sizes) and handles mirroring, striping, and other redundancy schemes for data reads/writes.

[0290] When the ZM needs to allocate a Zone, it will request an allocation of 2 or more sets of Regions. For example, a Zone may be allocated for 1 GB of data. The Regions that make up this Zone will be able to contain 1 GB of data including redundancy data. For a mirroring mechanism, the Zone will be made up of 2 sets of Regions of 1 GB each. Another example, a 3-disk striping mechanism utilize 3 sets of Regions of 1/2 GB each.

[0291] The ZM uses the ZR translation table (6) to find out the location (drive number and start Region number) of each set of Regions that makes up the Zone. Assuming a 1/2 GB Region size, a maximum of 24 Regions will be needed. 24 Regions make up 2x1 GB Zones. So the ZR translation table contains 24 columns that provide drive/region data.

[0292] The ZM works generally as follows:

[0293] a. In the case of SDM (single drive mirroring), 24 columns are used. The drive numbers are the same in all columns. Each entry corresponds to a physical Region on a physical drive that makes up the Zone. The first 12 entries point to Regions that contain one copy of the data. The last 12 entries point to the Regions containing the second copy of the data.

[0294] b. The case of DDM (dual drive mirroring) is the same as SDM except that the drive number on the first 12 entries is different from that in the last 12 entries.

[0295] c. In the case of striping, three or more columns may be used. For example, if striping is used across three drives, six Regions may be needed from three different drives (i.e., 18 entries are used), with the first six entries containing the same drive number, the next six entries containing another drive number, and the following six entries containing a third drive number; the unused entries are zeroed.

[0296] The following table shows a representation of the zone region translation table:

Zone#	Size of Zone	Size of each Re-gion	Us-age	Drive/Region (1)	Drive/Region (2)	...	Drive/Region (23)	Drive/Region (24)
0	1 GB	1/12	SDM	0, 2000	0, 1000	...	0, 10	0, 2000
1	1 GB	1/12	DDM	0, 8000	0, 3000	...	1, 2000	1, 10

-continued

Zone#	Size of Zone	Size of each Re-gion	Us-age	Drive/Region (1)	Drive/Region (2)	...	Drive/Region (23)	Drive/Region (24)
2	1 GB	1/12	SD3	3, 4000	3, 3000		4, 2000	4, 1000
...								
N			Free					

[0297] When a read/write request comes in, the ZM is provided with the Zone number and an offset into that Zone. The ZM looks in the ZR translation table to figure out the redundancy mechanism for that Zone and uses the offset to calculate which Drive/Region contains the sector that must be read/written. The Drive/Region information is then provided to the L1 layer to do the actual read/write. An additional possible entry in the Usage column is "Free". "Free" indicates that the Zone is defined but currently not used.

[0298] The cluster manager allocates and de-allocates clusters within the set of data Zones.

[0299] The Layout Manager provides run-time re-layout of the Zones vis-à-vis their Regions. This may occur as a result of disk insertion/removal or failure.

[0300] The Layer 1 (L1) software knows about physical drives and physical sectors. Among other things, the L1 software allocates Regions from physical drives for use by the Zones Manager. In this exemplary embodiment, each Region has a size of 1/2 GB (i.e., 174763 sectors) for a four-drive array system. A system with a larger maximum number of drives (8, 12 or 16) will have a different Region size.

[0301] In order to create a Zone containing 1 GB of data with SD3 (striping over three drives; two data plus parity), we would end up using six Regions each in three drives ($6 \times 1/2 = 1$ GB per drive).

[0302] The use of this Region scheme allows us to provide better utilization of disk space when Zones get moved around or reconfigured e.g., from mirroring to striping. The L1 software keeps track of available space on the physical drives with a bitmap of Regions. Each drive has one bitmap. Each Region is represented by two bits in the bitmap in order to track if the Region is free, used, or bad. When the L2 software (ZM) needs to create a Zone, it gets a set of Regions from the L1 layer. The Regions that make up a Zone are not contiguous within a disk.

[0303] Requests to L1 include:

[0304] a. Data read/write (to a cluster within a group of Regions)

[0305] b. Control data read/write (tables, data structures, DIC etc)

[0306] c. Allocate physical space for a Region (actual physical sectors within 1 drive)

[0307] d. De-allocate Region

[0308] e. Raw read/write to physical clusters within a physical drive

[0309] f. Copy data from one Region to another

[0310] g. Mark region as bad.

[0311] The free region bitmap may be large, and therefore searches to find the free entry (worst case is that no entries are free) may be slow. In order to improve performance, part of the bitmap can be preloaded into memory, and a linked list of free regions can be stored in memory. There is a list for each active zone. If a low water mark on the list is reached, more free entries can be read from the disk as a background activity.

[0312] The Disk Manager operates at layer 0. As shown in the following table, there are two sub-layers, specifically an abstraction layer and the device drivers that communicate with the physical storage array.

Layer 0a: Abstraction
Layer 0b: OS interface to device drivers and device drivers
Physical Storage Array Hardware

[0313] The Device Drivers layer may also contain several layers. For example, for a storage array using USB drives, there is an ATA or SCSI stack on top of the USB transport

layer. The abstraction layer provides basic read/write functions that are independent of the kinds of drives used in the storage array.

[0314] One or more disk access queues may be used to queue disk access requests. Disk access rates will be one of the key performance bottlenecks in our system. We will want to ensure that the disk interface is kept as busy as possible at all times so as to reduce general system latency and improve performance. Requests to the disk interface should have an asynchronous interface, with a callback handler to complete the operation when the disk operation has finished. Completion of one disk request will automatically initiate the next request on the queue. There may be one queue per drive or one queue for all drives.

[0315] Layer 1 will reference drives as logical drive numbers. Layer 0 will translate logical drive numbers to physical drive references (e.g., /dev/sda or file device number as a result of an open () call). For flexibility (expansion via USB), there should be a queue for each logical drive.

[0316] The following are some exemplary object definitions and data flows:

```

MSG object : incoming from host
    Lba
    Length
    LUN
    Data
REPLY object : outgoing to host
    Status
    Host
    Length
    Data
Data Read
Data read flow:
rc=lockm.islocked(MSG)
rc = catm.read(MSG, REPLY)
    status = zonem.read(zone, offset, length, buffer)
    regionm.read(logical_disk, region_number, region_offset, length, buffer)
    diskm.read(logical_disk, offset, length, buffer)
Data Write
Data write flow:
diskutildaemon.spaceavailable( )
journalm.write(MSG)
    lockm.lock(msg)
    zonem.write(journal_zone, offset, length, buffer)
    regionm.write          - journal entry
    diskm.write
    regionm.write          - end marker
    diskm.write
catm.write(MSG)
    catm.readcluster(lba, offset, length, buffer)    - if need to merge sector into cluster
    - merge
    "if(lba already allocated)"
    catm.readhashkey(lba)
    hashm.lookup(hashkey, zone, offset)
    "if(refcount==1)"
    hashentry.getrefcount( )
    hashm.remove(hashentry)
    hasm.add(shal, zone, offset)
    zonem.write(zone, offset, length, buffer)    - write data
    "else"
    hashentry.remove( )
    clusterm.allocate(zone, offset)    - allocate new cluster
    zonem.createzone(zone)
    regionm.unusedregions(logical_disk)
    regionm.allocate(logical_disk, number_regions, region_list)
    zonem.write(...)    - write data
    hashm.add(...)    - add new entry to hash table

```

-continued

```

“endif”
hashdaemon.add(lba,sha1)           - add to hash daemon Q
catm.writehashkey(lba,hashkey)      - copy new hash key to CAT
“else”
catm.update(lba,zone,offset,hashkey) - update CAT with new entry
“endif”
journalm.complete(MSG)
lockm.unlock(MSG)
    - update r/w cursors

```

[0317] The following is a description of physical disk layout. As discussed above, each disk is divided into Regions of fixed size. In this exemplary embodiment, each Region has a size of $\frac{1}{12}$ GB (i.e., 174763 sectors) for a four-drive array system. A system with a larger maximum number of drives (8, 12 or 16) will have a different Region size. Initially, Region numbers 0 and 1 are reserved for use by the Regions Manager and are not used for allocation. Region number 1 is a mirror of Region number 0. All internal data used by the Regions Manager for a given hard disk is stored in Region numbers 0 and 1 of this hard disk. This information is not duplicated (or mirrored) to other drives. If there are errors in either Region 0 or 1, other

Regions can be allocated to hold the data. The Disk Information Structure points to these Regions.

[0318] Each disk will contain a DIS that identifies the disk, the disk set to which it belongs, and the layout information for the disk. The first sector on the hard disk is reserved. The DIS is stored in the first non-bad cluster after the first sector. The DIS is contained in 1 KB worth of data. There are two copies of the DIS. The copies of the DIS will be stored on the disk to which it belongs. In addition, every disk in the system will contain a copy of all the DISs of the disks in the system. The following table shows the DIS format:

Offset	Size	Name	Value/Valid Range	Description
0	32 bytes	disStartSignature	“_DISC INFORMATION CLUSTER START_”	Identifies the cluster as being a possible disc information cluster. Cluster must be CRC'd to check that it is valid.
	WORD16	disVersion	Binary non-zero number	Identifies the structure version. This value is only changed when a material change is made to the structure layout or content meaning that makes it incompatible with previous versions of the Firmware.
	WORD16	disClusterSize	Binary non-zero number	The number of 512 byte sectors that make a cluster on this disc.
	WORD32	disCRC	CRC-32	CRC of the DIS structure.
	WORD32	disSize ^{!!!}		Size of DIS cluster (in bytes)
	WORD32	disDiskSet		The disk set this disk belongs to
	WORD32	disDriveNumber	0 to 15	The drive number within the disk set
	WORD32	disSystemUUID		UUID of the box this disk belongs to
	WORD64	disDiskSize		Size of the disk in number of sectors
	WORD32	disRegionSize		Size of Regions in number of sectors
	WORD64	disRegionsStart		Sector offset to the start of the first Region on the disk
	WORD64	disCopyOffset		Sector offset to where the copy of this DIS is stored. The disCopyOffset of each DIS reference each other

-continued

Offset	Size	Name	Value/Valid Range	Description
	WORD64	disDISBackup		Sector offset to the table containing the copies of the DISs of all the disks
	WORD32	disDISBackupSize		Number of DISs in the DIS Backup section
	WORD32	disRIS0Region		Region number of where first copy of the RIS is stored
	WORD32	disRIS0Offset		Number of sectors offset within the Region to the sector where the Regions Information Structure is located
	WORD32	disRIS1Region		For the copy of the RIS
	WORD32	disRIS1Offset		For the copy of the RIS
	WORD32	disZIS0Region		Region number of Region where the Zones Information Structure is located. This is ONLY used if there is a ZTR on this disk. Otherwise, it is zero.
	WORD32	disZIS0Offset		Offset to the ZIS within the region
	WORD32	disZIS1Region		Region number of Region where a copy of the ZIS is located. This is ONLY used in a single drive system. In other cases, this entry is 0.
	WORD32	disZIS1Offset		Offset to the ZIS within the region

[0319] Regions Manager stores its internal data in a regions information structure. The following table shows the regions information structure format:

[0320] The zones information structure provides information on where the Zones Manager can find the Zones Table. The following shows the zones information structure format:

Offset	Size	Name	Value/Valid Range	Description
0	WORD64	risSignature		Indicates that this is a RIS
	WORD32	risSize		Size of this structure (bytes)
	WORD32	risChecksum		Checksum
	WORD32	risVersion		Version of this table (and bitmap)
	WORD32	risDrive		Logical Drive number
	WORD64	risStartSector		Absolute start sector (in disk) of Regions utilization bitmap
	WORD32	risSectorOffset		Sector offset of Regions utilization bitmap within the current Region
	WORD32	risSizeBitmap		Size of bitmap (in bits?)
	WORD64	risNumberRegions		Number of regions on this disk (also implies size of bitmap)

Offset	Size	Name	Value/Valid Range	Description
0	WORD64	zisSignature		Indicates that this is a ZIS
8	WORD32	zisSize		Size of this structure (bytes)
12	WORD32	zisChecksum		Checksum
16	WORD32	zisVersion		Version of this table (and bitmap)
20	WORD16	zisFlags		Bit 0 = 1 if this disk is used to contain the Zones info Bits 14–15: redundancy type (either SDM or DDM only)
22	WORD16	zisOtherDrive		Logical drive number of the drive that contains the other copy of the Zones Table.
24	WORD32	zisNumberRegions		Number of Regions used to contain each copy of the Zones Table. Equal to the number of Zones Table Nodes.
28	WORD32	zisStartOffset		Byte offset pointing to start of linked list of Regions that are used to contain the Zones Table. Each entry in the linked list is called 'Zones Table Node'
	WORD32	zisNumberofZones		Number of Zones (entries in Zones Table) in the system
	WORD32	zisZoneSize		Size of zones in bytes

[0321] High level information zones contain the Zone tables and other tables used by the high level managers. These will be protected using mirroring.

[0322] The following table shows the zones table node format:

Size	Name	Description
WORD32	ztNextEntry	Pointer to next entry in linked list
WORD32	ztCount	Count of this entry
WORD64	ztRegion	Region number

[0323] The following is a description of layout of zones information. The linked list of Zones Table Nodes is placed after the ZIS in the following manner:

Zones Information Structure
First Zones Table Node (16bytes)
...
Last Zones Table Node (16 bytes)

[0324] This information is stored in the Zones Table Region.

[0325] FIG. 16 shows the drive layout in accordance with an exemplary embodiment of the invention. The first two regions are copies of one another. A third (optional) Zones Table Region contains the Zone Tables. In a system with more than one drive, only two of the drives contain a ZTR. In a system with only one drive, two Regions are used to hold the two (mirrored) copies of the ZTR. The DIS contains information on the location of the RIS and the ZIS. Note that the first copy of the RIS does not have to be in Region 0 (e.g., could be located in a different Region if Region 0 contains bad sectors).

[0326] The Zones Manager needs to load the Zones Tables on system start up. To do that, it extracts the Region number and offset from the DISs. This will point to the start of the ZIS.

[0327] Certain modules (e.g., the CAT Manager) store their control structures and data tables in Zones. All control structures for modules in Layer 3 and higher are referenced from structures that are stored in Zone 0. This means, for example, that the actual CAT (Cluster Allocation Tables) locations are referenced from the data structures stored in Zone 0.

[0328] The following table shows the zone 0 information table format:

Offset	Size	Name	Value/Valid Range	Description
0	WORD64	zitSignature		Indicates that this is a ZIT
	WORD32	zitSize		Size of this structure (bytes)
	WORD32	zitChecksum		Checksum of this structure
	WORD32	zitVersion		Version of this structure

-continued

Offset	Size	Name	Value/Valid Range	Description
	WORD32	zitCATLStartOffset		Byte offset (within this Zone) of start of CAT linked list
	WORD32	zitCATSize		Number of nodes in CAT linked list. Equal to number of Zones containing the CAT
	WORD64	zitCATAddressable		The max LBA supported by the CAT. Effectively the size of the CAT
	WORD32	zitHTStartOffset		Byte (within this Zone) of the start of the Hash Table linked list
	WORD32	zitHTNumberNodes		Number of nodes in Hash Table linked list
	WORD64	zitHTSize		Size of Hash Table data in bytes

[0329] The CAT linked list is a linked list of nodes describing the Zones that contain the CAT. The following table shows the CAT Linked List node format:

Size	Name	Description
WORD32	catllNextEntry	Pointer to next entry in linked list
WORD16	catllCount	Count of this entry
WORD16	catllZone	Zone number containing this portion of the CAT

[0330] The hash table linked list is a linked list of nodes that describe the Zones which hold the Hash Table. The following table shows the Hash Table Linked List node format:

Size	Name	Description
WORD32	htllNextEntry	Pointer to next entry in linked list
WORD16	htllCount	Count of this entry
WORD16	htllZone	Zone number containing this portion of the hash table

[0331] FIG. 17 demonstrates the layout of Zone 0 and how other zones are referenced, in accordance with an exemplary embodiment of the invention.

[0332] As discussed above, a Redundant set is a set of sectors/clusters that provides redundancy for a set of data. Backing up a Region involves copying the contents of a Region to another Region.

[0333] In the case of a data read error, the lower level software (Disk Manager or Device Driver) retries the read request two additional times after an initial failed attempt. The failure status is passed back up to the Zones Manager. The Zones Manager then attempts to reconstruct the data that is requested (by the read) from the redundant clusters in

the disk array. The redundant data can be either a mirrored cluster (for SDM, DDM) or a set of clusters including parity (for a striped implementation). The reconstructed data is then passed up back to the host. If the ZM is unable to reconstruct the data, then a read error is passed up back to the host. The Zones Manager sends an Error Notification Packet to the Error Manager. FIG. 18 demonstrates read error handling in accordance with an exemplary embodiment of the invention.

[0334] In the case of a data write error, the lower level software (Disk Manager or Device Driver) retries the write request two additional times after an initial failed attempt. The failure status is passed back up to the Zones Manager. The Zones Manager sends an Error Notification Packet to the Error Manager.

[0335] When a data write is performed at this level, the redundancy information is also written to disk. As a result, as long as only one cluster has a write error, a subsequent read will be able to reconstruct the data. If there are multiple disk errors and redundancy information cannot be read or written, then there are at least two possible approaches:

[0336] a. Return a write error status to the host. Back up all the Regions associated with the redundant set to newly allocated Regions that do not contain bad sectors.

[0337] b. Hold off the write. Back up all the Regions associated with the redundant set to newly allocated Regions that do not contain bad sectors. Subsequently, do the write on the appropriate cluster in the newly allocated Regions (along with all redundancy parts e.g., parity etc.). A separate write queue would be used to contain the writes that have been held off.

[0338] Approach (a) is problematic because a write status would likely have already been sent to the host as a result of a successful write of the Journal, so the host may not know that there has been an error. An alternative is to report a failure with a read, but allow a write. A bit in the CAT could be used to track that the particular LBA should return a bad read.

[0339] FIG. 19 demonstrates write error handling in accordance with an exemplary embodiment of the invention.

[0340] The Error Manager (EM) checks the cluster to see if it is really bad. If so, the entire region is considered bad.

The contents of the Region are copied over to a newly allocated Region on the same disk. The current Region is then marked BAD. While copying over the Region, the Error Manager will reconstruct data where necessary when it encounters bad sectors. **FIG. 20** is a logic flow diagram demonstrating backup of a bad Region by the Error Manager in accordance with an exemplary embodiment of the invention.

[0341] If there is a data read error and the Error Manager is unable to reconstruct the data for a given cluster (e.g., as a result of read errors across the redundant set) then zeros will be used in place of the data that cannot be reconstructed. In this case, other Regions (from the same Redundant Set) that contain bad sectors will also have to be backed up. Again, zeros will be used in place of the data that cannot be reconstructed.

[0342] Once a copy of the redundant set is made, the EM disables access to the clusters corresponding to this part of the Zone. It then updates the Zones Table to point to the newly allocated Regions. Subsequently, accesses to the clusters are re-enabled.

[0343] This exemplary embodiment is designed to support eight snapshots (which allows use of one byte to indicate whether hash/cluster entries are used by a particular snapshot instance). There are two tables involved with snapshots:

[0344] 1. A per-snapshot CAT table will need to exist to capture the relationship between logical sector addresses and the cluster on the disk that contains the data for that LSA. Ultimately the per-snapshot CAT must be a copy of the CAT at the moment the snapshot was taken.

[0345] 2. The system hash table, which maps between hash values and a data cluster. The hash function returns the same results regardless of which snapshot instance is being used, and as a result is common across all snapshots. As a result, this table must understand whether a unique cluster is being used by any snapshots. A hash cluster entry can not be freed, or replaced with new data unless there are no snapshots using the hash entry.

[0346] There will always be a snapshot that is current and being added to. When a hash entry is created or updated, we will need to apply the current snapshot number to that hash entry. When a snapshot is made, the current snapshot number will be incremented.

[0347] Clusters/hash entries that are not longer required by any snapshots are freed by walking through the hash table and find any hash entries with the retiring snapshot bit set and clearing that bit. If the snapshot byte is now zero, the hash entry can be removed from the table and the cluster can be freed.

[0348] To prevent collisions with any new entries being added to the hash tree (because the new snapshot number is the same as the retiring snapshot number), only allow 7 snapshots may be permitted to be taken, with the final (eighth) snapshot the one that is being retired. The hash table can be walked as a background activity.

[0349] In order to create a snapshot, a second CAT zone could be written whenever the main CAT is being updated. These updates could be queued and the shadow CAT could be updated as another task. In order to snapshot, the shadow CAT becomes the snapshot CAT.

[0350] Once the snapshot is done, a background process can be kicked off to copy this snapshot table to a new zone become the new snapshot CAT. A queue could be used so that the shadow CAT queue is not processed until the copy of the CAT had completed. If a failure were to occur before updating the shadow CAT (in which case entries in the queue may be lost), re-shadow from the primary CAT table could be performed before the array is brought online.

[0351] Alternatively, when a snapshot is required, a collection of "deltas" plus the initial CAT copy could make up the snapshot. A background task could then reconstitute a full snapshot CAT from this info. This would require little or no downtime to do the snapshot. In the meantime, another set of deltas could be collected for the following snapshot.

[0352] As discussed above, a so-called "garbage collector" may be used to free up clusters which are no longer used by the host file system (e.g., when a file is deleted). Generally speaking, garbage collection works by finding free blocks, computing their host LSAs, and locating their CAT entries based on the LSAs. If there is no CAT entry for a particular LSA, then the cluster is already free. If, however, the CAT entry is located, the reference count is decremented, and the cluster is freed if the count hits zero.

[0353] One issue with garbage collection is that it may be difficult to distinguish a block that the host filesystem has in use from one that it has previously used and at some point marked free. When the host filesystem writes a block, the storage system allocates a cluster for the data as well as a CAT entry to describe it. From that point on, the cluster will generally appear to be in use, even if the host filesystem subsequently ceases to use its block (i.e., the cluster will still be in use with a valid CAT entry).

[0354] For example, certain host filesystems use a bitmap to track its used disk blocks. Initially, the bitmap will indicate all blocks are free, for example, by having all bits clear. As the filesystem is used, the host filesystem will allocate blocks through use of its free block bitmap. The storage system will associate physical storage with these filesystem allocations by allocating clusters and CAT entries as outlined earlier. When the host filesystem releases some blocks back to its free pool, it simply needs to clear the corresponding bits in its free block bitmap. On the storage system, this will generally be manifested as a write to a cluster that happens to contain part of the host's free block bitmap, likely with no I/O to the actual cluster being freed itself (although there might be I/O to the freed cluster, for example, if the host filesystem were running in some enhanced security mode, in which case it would likely write zeros or a crypto strong hash of random data to the cluster in order to reduce the chance that stale cluster contents can be read by an attacker). Furthermore, there is no guarantee that the host filesystem will reuse blocks that it has previously freed when satisfying new allocation requests. Thus, if the host filesystem continues to allocate what from the storage system's point of view are new, i.e. previously unused, blocks then the storage system will quickly run out of free clusters, subject to whatever space can be reclaimed via compression. For example, assuming a filesystem block is 4k, if the host allocates filesystem blocks **100** through **500**, subsequently frees blocks **300** through **500**, and then allocates blocks **1000** through **1100**, the total filesystem usage will be 300 blocks, and yet the array will have 500 clusters in use.

[0355] In an exemplary embodiment of the present invention, the storage system may detect the release of host filesystem disk resources by accessing the host filesystem layout, parsing its free block bitmaps, and using that information to identify clusters that are no longer being used by the filesystem. In order for the storage system to be able to identify unused clusters in this way, the storage system must be able to locate and understand the free block bitmaps of the filesystem. Thus, the storage system will generally support a predetermined set of filesystems for which it “understands” the inner working sufficiently to locate and utilize the free block bitmaps. For unsupported filesystems, the storage system would likely be unable to perform garbage collection and should therefore only advertise the real physical size of the array in order to avoid being overcommitted.

[0356] In order to determine the filesystem type (e.g., NTFS, FAT, ReiserFS, ext3), the filesystem’s superblock (or an equivalent structure) needs to be located. To find the superblock, the partition table will be parsed in an attempt to locate the OS partition. Assuming the OS partition is located, the OS partition will be parsed in an attempt to locate the superblock and thereby identify the filesystem type. Once the filesystem type is known, the layout can be parsed to find the free block bitmaps.

[0357] In order to facilitate searching for free blocks, historical data of the host filesystem bitmap can be kept, for example, by making a copy of the free block bitmap that can be stored in a private, non-redundant zone and performing searches using the copy. Given the size of the bitmap, information may be kept for a relatively small number of clusters at a time rather than for the whole bitmap. When a garbage collection is performed, the current free block bitmap can be compared, cluster-by-cluster, with the historical copy. Any bitmap entries transitioning from allocated to free can be identified, allowing the scavenging operation to be accurately directed to clusters that are good candidates for reclamation. As each bitmap cluster is processed, the historical copy can be replaced with the current copy to maintain a rolling history of bitmap operations. Over time the copy of the free block bitmap will become a patchwork of temporally disjoint clusters, but since the current copy will always be used to locate free entries, this does not cause any problems.

[0358] Under certain conditions, there could be a race condition regarding the free block bitmap, for example, if the host filesystem allocates disk blocks using its free block bitmap, then writes its data blocks, then flushes the modified bitmap back to disk. In such a case, the garbage collector might free a cluster even though the filesystem is using the cluster. This could lead to filesystem corruption. The storage system should be implemented to avoid or handle such a condition.

[0359] Because garbage collection can be a fairly expensive operation, and since even lightweight scavenging will consume back-end I/O bandwidth, garbage collection should not be overused. The garbage collector should be able to run in several modes ranging from a light background lazy scavenge to an aggressive heavyweight or even high priority scavenge. For example, the garbage collector could be run lightly when 30% of space is used or once per week at a minimum, run slightly more heavily when 50% of space is

used, and run at a full high-priority scavenge when 90% or more of disk space is used. The aggressiveness of the garbage collector could be controlled by limiting it to a target number of clusters to reclaim and perhaps a maximum permissible I/O count for each collection run. For example, the garbage collector could be configured to reclaim 1 GB using no more than 10,000 I/Os. Failure to achieve the reclaim request could be used as feedback to the collector to operate more aggressively next time it is run. There may also be a “reclaim everything” mode that gives the garbage collector permission to parse the entire host filesystem free block bitmap and reclaim all blocks that it possibly can. This might be done as a last ditch attempt to reclaim clusters when the array is (almost) completely full. The garbage collector may be run periodically to apply its rules and may or may not decide to perform a scavenge operation. The scavenge operation should also be able to be explicitly requested from another module, for example the region manager when it is struggling to find clusters to build a region.

[0360] The garbage collection function can be tied into the status indicator mechanism. For example, at some point, the storage system might be in a “red” condition, although an ongoing garbage collection operation might free up enough space to erase the “red” condition. Additional indicator states could be employed to show related status information (e.g., the red indicator light might be made to blink to indicate that a garbage collection operation is ongoing).

[0361] FIG. 21 is a schematic block diagram showing the relevant components of a storage array in accordance with an exemplary embodiment of the present invention. Among other things, the storage array includes a chassis 2502 over which a storage manager 2504 communicates with a plurality of storage devices 2508₁-2508_N, which are coupled to the chassis respectively through a plurality of slots 2506₁-2506_N. Each slot 2506₁-2506_N may be associated with one or more indicators 2507₁-2507_N. Among other things, the storage manager 2504 typically includes various hardware and software components for implementing the functionality described above. Hardware components typically include a memory for storing such things as program code, data structures, and data as well as a microprocessor system for executing the program code.

Virtual Hot Spare

[0362] As discussed above, in many storage systems, a hot spare storage device will be maintained in a ready state so that it can be brought online quickly in the event another storage device fails. In certain embodiments of the present invention, rather than maintaining a physically separate hot spare, a virtual hot spare is created from unused storage capacity across a plurality of storage devices. Unlike a physical hot spare, this unused storage capacity is available if and when a storage device fails for storage of data recovered from the remaining storage device(s).

[0363] The virtual hot spare feature requires that enough space be available on the array to ensure that data can be re-laid out redundantly in the event of a disk failure. Thus, on an ongoing basis, the storage system typically determines the amount of unused storage capacity that would be required for implementation of a virtual hot spare (e.g., based on the number of storage devices, the capacities of the various storage devices, the amount of data stored, and the

manner in which the data is stored) and generates a signal if additional storage capacity is needed for a virtual hot spare (e.g., using green/yellow/red lights to indicate status and slot, substantially as described above). As zones are allocated, a record is kept of how many regions are required to re-layout that zone on a per disk basis. The following table demonstrates a virtual hot spare with four drives used:

Zone	Type	Stored		Regions required if disk fails			
		On Disks	Comments	Disk 0	Disk 1	Disk 2	Disk 3
2	Dual Drive Mirror	0, 1	Reconstruct on disk 2 or 3 if 0 or 1 fails	12	12	0	0
3	Dual Drive Mirror	0, 3	Reconstruct on disk 1 or 2 if 1 or 2 fails	12	0	0	12
5	Triple Drive Stripe	1, 2, 3	Reconstruct on disk 0 if 1, 2, or 3 fails	0	6	6	6
10	Four Drive Stripe	0, 1, 2, 3	Convert to triple drive strip across other three disks	2, 2, 2	2, 2, 2	2, 2, 2	2, 2, 2

[0364] The following table demonstrates a virtual hot spare with three drives used:

Zone	Type	Stored		Regions required if disk fails		
		On Disks	Comments	Disk 0	Disk 1	Disk 2
2	Dual Drive Mirror	0, 1	Reconstruct on disk 3	12	12	0
3	Dual Drive Mirror	0, 3	Reconstruct on disk 1	12	0	12
5	Triple Drive Stripe	1, 2, 3	Convert to dual drive mirror	6, 6	6, 6	6, 6

[0365] In this exemplary embodiment, virtual hot spare is not available on an array with only 1 or 2 drives. Based on the information for each zone and the number of disks in the array, the array determines a re-layout scenario for each possible disk failure and ensure that enough space is available on each drive for each scenario. The information generated can be fed back into the re-layout engine and the zone manager so that the data can be correctly balanced between the data storage and the hot spare feature. Note that the hot spare feature requires enough spare working space regions on top of those calculated from the zone layout data so that re-layout can occur.

[0366] FIG. 22 is a logic flow diagram showing exemplary logic for managing a virtual hot spare in accordance with an exemplary embodiment of the present invention. In block 2102, the logic determines a re-layout scenario for each possible disk failure. In block 2104, the logic determines the amount of space needed on each drive for re-layout of data redundantly in a worst case scenario. In block 2106, the logic determines the amount of spare working space regions needed for re-layout of data redundantly in a worst case scenario. In block 2108, the logic determines the total amount of space needed on each drive in order to permit re-layout of data redundantly in a worst case scenario

(essentially the sum of the amount of space needed for re-layout and the amount of spare working space regions needed). In block 2110, the logic determines whether the storage system contains an adequate amount of available storage. If there is an adequate amount of available storage (YES in block 2112), then the logic iteration terminates in block 2199. If, however, there is an inadequate amount of

available storage (NO in block 2112), then the logic determines which drive/slot requires upgrade, in block 2114. Then, in block 2116, the logic signals that additional storage space is needed and indicates which drive/slot requires upgrade. The logic iteration terminates in block 2199.

[0367] FIG. 23 is a logic flow diagram showing exemplary logic for determining a re-layout scenario for each possible disk failure, as in block 2102 of FIG. 22, in accordance with an exemplary embodiment of the present invention. In block 2202, the logic allocates a zone. Then, in block 2204, the logic determines how many regions are required to re-layout that zone on a per-disk basis. The logic iteration terminates in block 2299.

[0368] FIG. 24 is a logic flow diagram showing exemplary logic for invoking the virtual hot spare functionality in accordance with an exemplary embodiment of the present invention. In block 2302, the logic maintains a sufficient amount of available storage to permit re-layout of data redundantly in the event of a worst case scenario. Upon determining loss of a drive (e.g., removal or failure), in block 2304, the logic automatically reconfigures the one or more remaining drives to restore fault tolerance for the data, in block 2306. The logic iteration terminates in block 2399.

[0369] FIG. 25 is a logic flow diagram showing exemplary logic for automatically reconfiguring the one or more remaining drives to restore fault tolerance for the data, as in block 2306 of FIG. 24, in accordance with an exemplary embodiment of the present invention. In block 2402, the logic may convert a first striped pattern across four or more storage devices to a second striped pattern across three or more remaining storage devices. In block 2404, the logic may convert a striped pattern across three storage devices to a mirrored pattern across two remaining storage devices. Of course, the logic may convert patterns in other ways in order to re-layout the data redundantly following loss of a drive. The logic iteration terminates in block 2499.

[0370] With reference again to FIG. 21, the storage manager 2504 typically includes appropriate components and logic for implementing the virtual hot spare functionality as described above.

Dynamic Upgrade

[0371] The logic described above for handling dynamic expansion and contraction of storage can be extended to provide a dynamically upgradeable storage system in which storage devices can be replaced with a larger storage devices as needed, and existing data is automatically reconfigured across the storage devices in such a way that redundancy is maintained or enhanced and the additional storage space provided by the larger storage devices will be included in the pool of available storage space across the plurality of storage devices. Thus, when a smaller storage device is replaced by a larger storage device, the additional storage space can be used to improve redundancy for already stored data as well as to store additional data. Whenever more storage space is needed, an appropriate signal is provided to the user (e.g., using green/yellow/red lights substantially as described above), and the user can simply remove a storage device and replace it with a larger storage device.

[0372] FIG. 26 is a logic flow diagram showing exemplary logic for upgrading a storage device, in accordance with an exemplary embodiment of the present invention. In block 2602, the logic stores data on a first storage device in a manner that the data stored thereon appears redundantly on other storage devices. In block 2604, the logic detects replacement of the first storage device with a replacement device having greater storage capacity than the first storage device. In block 2606, the logic automatically reproduces the data that was stored on the first device onto the replacement device using the data stored redundantly on other devices. In block 2608, the logic makes the additional storage space on the replacement device available for storing new data redundantly. In block 2610, the logic may store new data redundantly within the additional storage space on the replacement device if no other device has a sufficient amount of available storage capacity to provide redundancy for the new data. In block 2612, the logic may store new data redundantly across multiple storage devices if at least one other device has a sufficient amount of available storage capacity to provide redundancy for the new data.

[0373] With reference again to FIG. 21, the storage manager 2504 typically includes appropriate components and logic for implementing the dynamic upgrade functionality as described above.

Miscellaneous

[0374] Embodiments of the present invention may be employed to provide storage capacity to a host computer, e.g., using a peripheral connect protocol in the manner described in my U.S. Provisional Patent Application No. 60/625,495, which was filed on Nov. 5, 2004 in the name of Geoffrey S. Barrall, and is hereby incorporated herein by reference in its entirety.

[0375] It should be noted that a hash algorithm may not produce hash values that are strictly unique. Thus, is it conceivable for the hash algorithm to generate the same hash value for two chunks of data having non-identical content. The hash function (which generally incorporates the hash algorithm) typically includes a mechanism for confirming uniqueness. For example, in an exemplary embodiment of the invention as described above, if the hash value for one chunk is different than the hash value of another chunk, then the content of those chunks are considered to be non-

identical. If, however, the hash value for one chunk is the same as the hash value of another chunk, then the hash function might compare the contents of the two chunks or utilize some other mechanism (e.g., a different hash function) to determine whether the contents are identical or non-identical.

[0376] It should be noted that the logic flow diagrams are used herein to demonstrate various aspects of the invention, and should not be construed to limit the present invention to any particular logic flow or logic implementation. The described logic may be partitioned into different logic blocks (e.g., programs, modules, functions, or subroutines) without changing the overall results or otherwise departing from the true scope of the invention. Often times, logic elements may be added, modified, omitted, performed in a different order, or implemented using different logic constructs (e.g., logic gates, looping primitives, conditional logic, and other logic constructs) without changing the overall results or otherwise departing from the true scope of the invention.

[0377] The present invention may be embodied in many different forms, including, but in no way limited to, computer program logic for use with a processor (e.g., a microprocessor, microcontroller, digital signal processor, or general purpose computer), programmable logic for use with a programmable logic device (e.g., a Field Programmable Gate Array (FPGA) or other PLD), discrete components, integrated circuitry (e.g., an Application Specific Integrated Circuit (ASIC)), or any other means including any combination thereof.

[0378] Computer program logic implementing all or part of the functionality previously described herein may be embodied in various forms, including, but in no way limited to, a source code form, a computer executable form, and various intermediate forms (e.g., forms generated by an assembler, compiler, linker, or locator). Source code may include a series of computer program instructions implemented in any of various programming languages (e.g., an object code, an assembly language, or a high-level language such as Fortran, C, C++, JAVA, or HTML) for use with various operating systems or operating environments. The source code may define and use various data structures and communication messages. The source code may be in a computer executable form (e.g., via an interpreter), or the source code may be converted (e.g., via a translator, assembler, or compiler) into a computer executable form.

[0379] The computer program may be fixed in any form (e.g., source code form, computer executable form, or an intermediate form) either permanently or transitorily in a tangible storage medium, such as a semiconductor memory device (e.g., a RAM, ROM, PROM, EEPROM, or Flash-Programmable RAM), a magnetic memory device (e.g., a diskette or fixed disk), an optical memory device (e.g., a CD-ROM), a PC card (e.g., PCMCIA card), or other memory device. The computer program may be fixed in any form in a signal that is transmittable to a computer using any of various communication technologies, including, but in no way limited to, analog technologies, digital technologies, optical technologies, wireless technologies (e.g., Bluetooth), networking technologies, and internetworking technologies. The computer program may be distributed in any form as a removable storage medium with accompanying printed or electronic documentation (e.g., shrink wrapped software),

preloaded with a computer system (e.g., on system ROM or fixed disk), or distributed from a server or electronic bulletin board over the communication system (e.g., the Internet or World Wide Web).

[0380] Hardware logic (including programmable logic for use with a programmable logic device) implementing all or part of the functionality previously described herein may be designed using traditional manual methods, or may be designed, captured, simulated, or documented electronically using various tools, such as Computer Aided Design (CAD), a hardware description language (e.g., VHDL or AHDL), or a PLD programming language (e.g., PALASM, ABEL, or CUPL).

[0381] Programmable logic may be fixed either permanently or transitorily in a tangible storage medium, such as a semiconductor memory device (e.g., a RAM, ROM, PROM, EEPROM, or Flash-Programmable RAM), a magnetic memory device (e.g., a diskette or fixed disk), an optical memory device (e.g., a CD-ROM), or other memory device. The programmable logic may be fixed in a signal that is transmittable to a computer using any of various communication technologies, including, but in no way limited to, analog technologies, digital technologies, optical technologies, wireless technologies (e.g., Bluetooth), networking technologies, and internetworking technologies. The programmable logic may be distributed as a removable storage medium with accompanying printed or electronic documentation (e.g., shrink wrapped software), preloaded with a computer system (e.g., on system ROM or fixed disk), or distributed from a server or electronic bulletin board over the communication system (e.g., the Internet or World Wide Web).

[0382] This application is related to the following U.S. Patent Applications, which are being filed on even date herewith and are hereby incorporated herein by reference in their entireties:

[0383] Attorney Docket No. 2950/104 entitled Dynamically Upgradeable Fault-Tolerant Storage System Permitting Various Sized Storage Devices and Method;

[0384] Attorney Docket No. 2950/105 entitled Dynamically Expandable and Contractible Fault-Tolerant Storage System With Virtual Hot Spare; and

[0385] Attorney Docket No. 2950/107 entitled Storage System Condition Indicator and Method.

[0386] The present invention may be embodied in other specific forms without departing from the true scope of the invention. The described embodiments are to be considered in all respects only as illustrative and not restrictive.

What is claimed is:

1. A method of storing data in a set of storage devices, the set having at least one storage device, the method comprising:

automatically determining a mixture of redundancy schemes, from among a plurality of redundancy schemes, for storing the data in the set of storage devices in order to provide fault tolerance and, when the set of storage devices happens to include a plurality of devices of different storage capacities, enhanced storage efficiency; and

storing the data in the set of storage devices using the mixture of redundancy schemes.

2. A method according to claim 1, further comprising:

automatically reconfiguring data previously stored on a first arrangement of storage devices using a first redundancy scheme to being stored using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices for accommodating without data loss at least one of expansion of capacity by the addition of another storage device to the set and contraction of capacity by the removal of a storage device from the set.

3. A method of storing data in a set of storage devices in a manner permitting dynamic expansion and contraction of the set, the set having at least one storage device, the method comprising:

storing data on the set of storage devices using a first redundancy scheme; and

upon a change in the composition of the set of storage devices, automatically reconfiguring the data on the set of storage devices using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices.

4. A method according to any of claims 1, 2, or 3, wherein the redundancy schemes are selected from the group including mirroring, striping with parity, RAID6, dual parity, diagonal Parity, Low Density Parity Check codes, and turbo codes.

5. A method according to any of claims 2 or 3, wherein reconfiguring the data comprises:

converting a mirrored pattern across two storage devices to a mirrored pattern on a single storage device.

6. A method according to any of claims 2 or 3, wherein reconfiguring the data comprises:

converting a mirrored pattern across a first pair of storage devices to a mirrored pattern across a second pair of storage devices.

7. A method according to any of claims 2 or 3, wherein reconfiguring the data comprises:

converting a striped pattern across three storage devices to a mirrored pattern across two storage devices.

8. A method according to any of claims 2 or 3, wherein reconfiguring the data comprises:

converting a first striped pattern across a first plurality of storage devices to a second striped pattern across a second plurality of storage devices.

9. A method according to any of claims 2 or 3, wherein reconfiguring the data comprises:

converting a mirrored pattern across one storage device to a mirrored pattern across two storage devices.

10. A method according to any of claims 2 or 3, wherein reconfiguring the data comprises:

converting a mirrored pattern on one storage device to a mirrored pattern on another storage device.

11. A method according to any of claims **2** or **3**, wherein reconfiguring the data comprises:

converting a mirrored pattern across two storage devices to striped pattern across three storage devices.

12. A method according to any of claims **2** or **3**, wherein reconfiguring the data comprises:

converting a first striped pattern across a first plurality of storage devices to a second striped pattern across a second plurality of storage devices including a storage device added to the set.

13. A method according to any of claims **1**, **2**, or **3**, further comprising:

parsing an object into chunks, each chunk having particular content; and

storing in the storage system only chunks having unique content.

14. A method according to claim **13**, wherein storing only chunks having unique content comprises:

storing in the storage system only chunks identified as having unique content based on a predetermined hash function.

15. A method according to claim **14**, wherein storing only chunks identified as being unique based upon a predetermined hash function comprises:

computing a hash number for each chunk; and

storing in the storage system a unique identifier for the object linked to hash numbers for the stored chunks and location information therefor.

16. A method according to any of claims **15**, wherein storing a unique identifier for the object includes using an object table that lists the unique identifier for the object and the hash numbers for the stored chunks.

17. A method according to any of claims **15**, wherein storing location information includes storing in a chunk table a listing for each chunk stored therein including a representation of its physical location in the storage system.

18. A method according any of claims **13**, further comprising:

translating objects stored in the storage system into a representation of at least one of a single storage device and a file system and an object storage system.

19. A method according to any of claims **13**, further comprising:

managing physical allocation of storage of chunks in the storage system so as to provide fault-tolerant storage, even when the storage system may include storage devices of varying capacity.

20. A method according to any of claims **15**, further comprising:

upon retrieving a particular chunk from the storage system, verifying the contents of the retrieved chunk based on the stored hash number for the chunk.

21. A method according to claim **20**, wherein verifying the contents of the retrieved chunk based on the stored hash number for the chunk comprises:

computing a hash number for the retrieved chunk; and

comparing the computed hash number with the corresponding stored hash number for the chunk.

22. A method of storing data in a set of storage devices, the method comprising:

storing data on the set of storage devices using at least one redundancy scheme;

detecting removal of a storage device from the set;

reconfiguring a portion of the data from the removed storage device using the remaining storage devices;

detecting reinsertion of the removed storage device into the set;

freeing regions of the reinserted storage device corresponding to said portion; and

refreshing any remaining regions of the reinserted storage device that may have been modified between removal and reinsertion.

23. A system for storing data, the system comprising:

a set of storage devices, the set having at least one storage device; and

a storage manager for storing the data in the set of storage devices using a mixture of redundancy schemes, the storage manager automatically determining the mixture of redundancy schemes, from among a plurality of redundancy schemes, for storing the data in the set of storage devices in order to provide fault tolerance and, when the set of storage devices happens to include a plurality of devices of different storage capacities, enhanced storage efficiency.

24. A system according to claim **23**, wherein the storage manager automatically reconfigures data previously stored on a first arrangement of storage devices using a first redundancy scheme to being stored using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices for accommodating without data loss at least one of expansion of capacity by the addition of another storage device to the set, and the contraction of capacity by the removal of a storage device from the set.

25. A system for storing data, the system comprising:

a set of storage devices, the set having at least one storage device; and

a storage manager for storing the data in the set of storage devices, the storage manager storing data on the set of storage devices using a first redundancy scheme and, upon a change in the composition of the set of storage devices, automatically reconfiguring the data on the set of storage devices using one of (a) a second redundancy scheme on the same arrangement of storage devices, (b) a second redundancy scheme on a different arrangement of storage devices, and (c) the same redundancy scheme on a different arrangement of storage devices.

26. A system according to any of claims **23**, **24**, or **25**, wherein the redundancy schemes are selected from the group including mirroring, striping with parity, RAID6, dual parity, diagonal Parity, Low Density Parity Check codes, and turbo codes.

27. A system according to any of claims **24** or **25**, wherein the storage manager converts a mirrored pattern across two storage devices to a mirrored pattern on a single storage device.

28. A system according to any of claims **24** or **25**, wherein the storage manager converts a mirrored pattern across a first pair of storage devices to a mirrored pattern across a second pair of storage devices.

29. A system according to any of claims **24** or **25**, wherein the storage manager converts a striped pattern across three storage devices to a mirrored pattern across two storage devices.

30. A system according to any of claims **24** or **25**, wherein the storage manager converts a first striped pattern across a first plurality of storage devices to a second striped pattern across a second plurality of storage devices.

31. A system according to any of claims **24** or **25**, wherein the storage manager converts a mirrored pattern across one storage device to a mirrored pattern across two storage devices.

32. A system according to any of claims **24** or **25**, wherein the storage manager converts a mirrored pattern on one storage device to a mirrored pattern on another storage device.

33. A system according to any of claims **24** or **25**, wherein the storage manager converts a mirrored pattern across two storage devices to striped pattern across three storage devices.

34. A system according to any of claims **24** or **25**, wherein the storage manager converts a first striped pattern across a first plurality of storage devices to a second striped pattern across a second plurality of storage devices including a storage device added to the set.

35. A system according to any of claims **23**, **24**, or **25**, further comprising:

a chunk parser, for parsing an object into chunks, each chunk having particular content, wherein the storage manager stores in the set of storage devices only chunks having unique content.

36. A system according to claim **35**, further comprising:

a hash code generator for generating a hash number for each chunk, the hash code generator and the physical storage manager in communication with an object table store and a chunk table store so that the object table lists a unique identifier for the object linked to hash numbers for the stored chunks and the chunk table lists for each chunk stored therein a representation of its physical location in the set of storage devices.

37. A system according to claim **35**, further comprising:

a fault tolerant manager, in communication with the storage manager, for managing physical allocation of storage of chunks in the set of storage devices so as to

provide fault-tolerant storage, even when the set of storage devices may include storage devices of varying capacity.

38. A system according any of claims **36**, wherein, upon retrieving a particular chunk from the set of storage devices, the physical storage manager verifies the contents of the retrieved chunk based on the stored hash number for the chunk.

39. A system according to claim **38**, wherein the physical storage manager computes a hash number for the retrieved chunk and compares the computed hash number with the corresponding stored hash number for the chunk.

40. A method for freeing unused storage blocks in a storage system, the method comprising:

identifying a free block;

determining a logical sector address for the free block;

identifying a cluster access table entry for the free block based on the logical sector address;

decrementing a reference count associated with the cluster access table entry; and

freeing the block if the decremented reference count is zero.

41. A method according to claim **40**, wherein identifying a free block comprises:

identifying a free block bitmap associated with the host filesystem; and

parsing the free block bitmap to identify clusters that are no longer being used by the filesystem.

42. A method according to claim **41**, wherein identifying the free block bitmap associated with the host filesystem comprises:

maintaining a partition table;

parsing the partition table to locate an operating system partition;

parsing the operating system partition to locate a superblock; and

identifying the host filesystem based on the superblock.

43. A method according to claim **41**, further comprising:

making a working copy of the free block bitmap, wherein the working copy of the free block bitmap is parsed to identify clusters that are no longer being used by the filesystem.

* * * * *